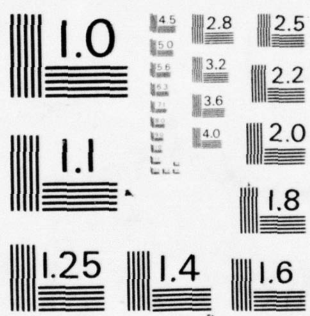
AD-A032 378 NAVAL POSTGRADUATE SCHOOL MONTEREY CALIF F/G 5/9
METHODOLOGIES OF OFFICER BILLET CLASSIFICATION.(U)
SEP 76 J LEMKE

UNCLASSIFIED NL

1 OF 1
AD
A032378

END
DATE
FILMED
1-77



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

2

NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

METHODOLOGIES OF
OFFICER BILLET CLASSIFICATION

by

Juergen Lemke

September 1976

Thesis Advisor: R. R. Read

Approved for public release; distribution unlimited.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (9) Methodologies of Officer Billet Classification,		5. TYPE OF REPORT & PERIOD COVERED (9) Master's Thesis; September 1976
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Juergen Lemke		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		(11) 12. REPORT DATE September 1976
		13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Naval Postgraduate School Monterey, California 93940	(12) 92p.	15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Cluster Analysis
Numerical Classification
Job Classification
Officer Billet Classification

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Four potentially valuable methods to classify officer billets into subgroups on the basis of multivariate observations about the billets are presented. The methods aiming to reduce the dimensionality and to identify homogeneous subgroups are: Principal Component Analysis, Multidimensional Scaling, Hierarchical Cluster Analysis (Hiclust) and Cluster Analysis optimizing an objective function (K-Means). They are (cont on p 2)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
(Page 1) S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

251 450

1/B

(cont fr p. 1)

applied to a data set obtained from an outside source and comprising 96 Navy officer billets. Thirteen quantitative variables measuring the relative amount of time spent for managerial responsibilities and resources have been entered into the analysis. On the basis of the entered variables, the presence of eight billet clusters have been determined. The evolved groups are described by their centroids and within group standard deviations.

Methodologies of Officer Billet Classification

by

Juergen Lemke
Major, Federal German Air Force
Ing. Grad., Technische Akademie der Luftwaffe Neubiberg, 1968

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL

September 1976

Author: ____________________

Approved by: ____________________
Thesis Advisor

____
Second Reader

____
Chairman, Department of Operations Research

____
Academic Dean

ABSTRACT

Four potentially valuable methods to classify officer billets into subgroups on the basis of multivariate observations about the billets are presented. The methods aiming to reduce the dimensionality and to identify homogeneous subgroups are: Principal Component Analysis, Multidimensional Scaling, Hierarchical Cluster nalysis (Hiclust) and Cluster Analysis Optimizing an Objective Function (K-Means). They are applied to a data set obtained from an outside source and comprising 96 Navy officer billets. Thirteen quantitative variables measuring the relative amount of time spent for managerial responsibilities and resources have been entered into the analysis. On the basis of the entered variables, the presence of eight billet clusters have been determined. The evolved groups are described by their centroids and within-group standard deviations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

A. SUBJECT AND PURPOSE

This thesis constitutes a first cut towards classifying Navy officer billets by an investigation of some methods which are of potential value to accomplish this classification. Clearly, every grouping has to be based on some information about the objects. Billets can be described by any set of variables measured or determined on them. Among such possible descriptors are the activities that are performed, the human behavior that is involved in the work activities, and the human qualities that are relevant for successful performance. It should be clear that in general a different set of descriptors will result in a different classification, and the most difficult task is to decide on the "appropriate" set of descriptors. Descriptors, or variables, are appropriate if they reflect the purpose of the classification and if they allow a significant grouping to take place. Satisfying both conditions is not as trivial as it may appear. In most cases, the judgment about the significance can be made only by applying the cross-validation technique because robust tests seem not to be available.

The variables of concern for this study are those derived from management theory. A questionnaire included as Appendix A, and developed by Professor Elster and Professor Read, served as the data collecting tool. The items in the questionnaire

are designed to measure the relative amount of time the officers spend in performing management functions and handling certain resources. These times have been recorded for various navy officer billets.

The solution of the variable selection problem, as stated above, is not the goal of this study despite its paramount importance. The adequacy of the items comprising the questionnaire will be assessed with respect to the significance of the evolving groups as a by-product. The data collected by the questionnaire serves as a coarse base to which the methods will be applied. It is not claimed that the data, comprising 96 samples, is representative for the whole population of Navy officer billets. Thus, the resulting classification is exploratory in nature.

The main purpose is to describe some numerical methods which are potential candidates to classify billets into homogeneous subgroups on the basis of qualitative and quantitative variables. The methods investigated and applied herein are:

- Principal Component Analysis
- Multidimensional Scaling
- Hierarchical Cluster Analysis (Hiclust)
- Cluster Analysis which Optimizes an Objective Function (K-Means).

The methods will be contrasted and their usefulness for the problem at hand will be discussed. Finally, the results of their application will be presented.

The long range goal is to develop a billet classification, sufficiently detailed so that educational and training requirements can be clearly specified. It is conceivable that such

detailed classification of Navy officer billets -- possibly one that resembles hierarchical structures known from biology -- is potentially beneficial, both for the Navy and for the individual serving in the Navy. Other areas of impact could be coding of billet, compensation and assignment. The more immediate goal is the development of a data-gathering instrument.

B. BACKGROUND

In the past, related work has been done by Stogdill and Shartle [13] in 1955. They analyzed the job responsibilities of 470 Navy officers and 66 business executives. A classification of the assignments has not been attempted. They have used the relative amount of time spent for 14 major management responsibilities as one set of items. About half of those responsibilities have been identified by early management theorists such as Fayol, Gaulic and several others at the beginning of the century.

It is noted in passing that Hemphill [6] identified quite different dimensions in an empirical research. He factor-analyzed 191 characteristics of 93 business executives.

Mahoney, Jerdee and Carrol [10], in their study on 452 managerial jobs, have first made the division between what they called "functional dimensions of management performance" and "subject dimensions of management performance". The "functional dimensions" are planning, investigating, coordinating, supervising, evaluating, staffing, negotiating and representing. The "subject dimensions" are employees, money

and finances, materials and goods, purchases and sales, methods and procedures, and facilities and equipment. They cross-cut the functional dimensions and were introduced to separate between the areas of competence. The study also established a classification by identifying "patterns of performance". Yet the philosophy of their grouping is quite different from the approach undertaken in this thesis. They preconceived the job groups solely from the functional dimensions without any look on the data. Instead of being generated from the data, the groups have been fixed in advance to be Planner, Investigator, Coordinator, Evaluator, Supervisor, Negotiator, Multispecialist, and Generalist.

The executive jobs have then been appointed to the groups mainly according to the high scores on the associated dimension. This procedure makes use of an external group model and resembles in some way the bayesian schemes. The subject dimensions were not entered into the classification. Its scores were merely used to assess the averages on these dimensions within the different groups.

The division between the functional dimensions and the subject dimensions reappears in the current questionnaire. Its first set of items is identical to the functional dimensions. The second set, called "Resources and Subject Matter", corresponds to the subject dimensions of the cited study. It has been adjusted to fit the military areas of competence.

Methodologies of classifying jobs more on the lines of adapting internal group models have been applied by De Nisi

and McCormick [1] in 1974. They recorded 187 job elements on 3700 jobs by a "Position Analysis Questionnaire" (PAQ) and performed a principal component analysis. The jobs then have been clustered by two clustering algorithms according to the component scores. The algorithms have identified 33 and 45 job classes respectively.

C. GENERAL SUMMARY OF CONCLUSIONS

The essential conclusions drawn from this study concerning the method and the resulting grouping are as follows:

1. Sophisticated numerical classification methods often require the representation of the data as points in a low dimensional euclidean space. Such representations may be achieved by applying either Principal Component Analysis or Multidimensional Scaling. Both methods point to about five basic dimensions. Grouping the billets according to those low dimensional configurations appears to be too coarse.

2. The Hierarchical Clustering Program (Hiclust) does not require strong metric assumptions. Its results are only satisfying if the data contains well-pronounced subgroups. The billet data does not seem to possess this desirable property so that the application of Hiclust does not promise useful results. More elaborate numerical classification may be achieved by applying an iterative clustering algorithm (K-Means). It is the most viable cluster method for the problems of classifying officer billets. Eight clusters have been identified by K-Means in this pilot study. They are

rather sensitive to changes of the data. The cluster solution is displayed in Appendix D.

3. Billet titles appear to be of little use for classification purposes when managerial responsibility and resource items are the input characteristics of the classification analysis. There seems to be greater homogeneity of responsibilities existing within the identified classes than within any common grouping by billet title. Even billets with the same title, but from different commands and locations, may exhibit rather different profile scores.

4. The discriminating ability of the currently used items vary widely between the identified clusters. An analysis of the overall within-cluster variation indicates that the resource items are superior if compared with the responsibility items.

5. The next developmental step should apply these methods of analysis to a more representative sample which should be large so that the cross-validation technique could be applied. The current instrument appears to be competently designed. Its main area of application should be limited to high ranking officers because junior officers often have to fulfill specific tasks. These tasks are difficult to describe accurately by the responsibility items of the questionnaire. This is believed to be the main reason that the current data is lacking the desirable structure.

The report will be organized as follows. Chapter II illuminates the data base and documents the treatment of outliers.

The conclusions concerning the data collecting tool based on the direct observation of the data are given. The presentation and application of Principal Component Analysis and Multidimensional Scaling are contained in Chapter III. Chapter IV describes the actual classification methods, Hierarchical Cluster Analysis (Hiclust) and Cluster Analysis Optimizing an Objective Function (K-Means) and their application to the billet data. The choices of the various parameter specifications are discussed with respect to the problem at hand. In the last chapter, a summary concerning the methodologies and the resulting classification is presented. Details are documented in the Appendices.

II. NATURE OF THE DATA BASE

The data base for the study comprises 96 data profiles obtained from Navy officers assigned to the Naval Postgraduate School. The data has been collected in summer 1975 by means of the questionnaire included as Appendix A. The respondents have been asked to describe their last non-student, non-operational billet by allocating the relative amount of time they had spent for the managerial activities and resources discussed above. In addition, the questionnaire provides for information about the importance the respondents felt that each of the first eight activities have had to successfully perform in the described billet. The three sets of items will be referred to (from now on) as responsibility items, responsibility importance items and resource items. Optional "other" items could be specified if the respondent felt that his billet had been concerned with responsibilities and resources not covered by the items in the questionnaire. Valuable information has been gained by this provision for the development of instruments for future studies. On the other hand, this feature has made the classification goal more complicated. Nineteen respondents have allocated time to self-defined items, with ten of them specifying 30% or more of their time.

Allocation of time to self-defined items had the effect of lowering the fixed items because of the percentage scale.

It was observed nearly always that a group evolved, having only one feature in common: All its members had assigned a significant proportion of time to self-defined items. It became apparent that the data profiles had been dampened by a substantial amount as a result.

For our purposes, such grouping is deemed invalid. Choices available for circumventing the difficulty include a) creating of additional items, and b) eliminating sample units which scored 30% or more on self-defined items. The former choice (a) would enlarge the number of items by about 15. The latter (b) would decrease the sample size by 10. Both of the consequences were regarded as distasteful.

It was decided to perform the following manipulations. Observations 22, 47 and 58 were discarded from the analysis. The corresponding billet names are missing and all of these scored between 47% and 50% on self-defined items. The other four most seriously distorted sample units are 20, 55, 69 and 95. Their scores were reallocated from self-defined to original items using the expertise of LCDR Wicker, a former detailer of junior officers.

Twelve respondents did not insert the title of their billets. Another seven failed to include the set of resource items. There was no possibility to recover the missing billet names because the survey was conducted anonymously. However, all those who did not respond to the set of resource items have given their billet titles. The missing profile entries 15, 33, 37, 75, 76, 78 and 94 were also estimated by

LCDR Wicker. This procedure was cross-validated. Using several subjects whose resource items were now missing, LCDR Wicker obtained a very satisfactory agreement with the actual allocation. The raw data and the described manipulations are documented in Appendix B.

One is naturally concerned with the question of reconciling this data salvaging procedure with the conclusion that billet titles do not appear to be useful discriminators for our data. Consequently, the analysis was repeated using only 82 vectors, which excludes those salvaged. The results are in Appendix E. Briefly, this exercise showed that although the cluster solution is not reproduced quite as well as had been hoped, the conclusion concerning the usefulness of billet titles remains.

The set of responsibility importance items has been excluded from the analysis. As might be expected, the correlation between respective items from the responsibility set and the responsibility importance set is very high. The product moment correlations were computed, the smallest yielding a value of .6058, which is significant at the 5% level (one-sided alternative). Thus, the inclusion of the responsibility importance items would lead to a greater implicit weight for the eight responsibility items in contrast to the resource items.

The net sample size of 93 seems to be small and a stable grouping is not expected. Also, the sample is unlikely to be representative of all billets in the Navy. The fact that

the respondents have been assigned to a postgraduate education has certainly influenced the composition of the sample. Yet, as a base for this pilot study, the sample is believed to be sufficient with regard to the stated goals. Major statistical inferences have not been planned. Similarly, stratification by subsets such as ranks has not been undertaken.

For the purpose of the classification, it has been decided to treat different responses with identical billet titles independently from one another and not to average the profiles. This allows determination of whether billets with the same title are classified into one group or scattered into several different groups. It will be seen that the numerical classification gives examples for both. Some billets with more than one response appear in one and only one group, as is the case for all four Detailers and two Ship Superintendents, Naval Shipyard. Other billets, such as Instructors, or Flag Lieutenants and Aides, emerge in more than one cluster.

There are three reasons which might be cited to account for this.

1. The reliability of the estimates over time are expected to be low.

2. The complexity of the items lead to different perceptions of what planning, evaluating or handling resources like Methods and Procedures is.

3. The incumbents or their superiors are likely to be the main determinators for the responsibilities and resources encountered, not the requirements of the billets.

It is obvious that the reliability of the estimates over time cannot be investigated from the data of this study.

It would require conducting the same survey on the same officers again after a certain time has passed. In Ref.[10] reliability results have been reported for the responsibility items and very similar resource items. After three weeks, 76% of their responses, describing the current executive job, have varied no more than 5% from the original estimates on the responsibility items. The corresponding reliability of the resource items has been found to be 70%. Taking into account that the billets described by our respondents were occupied by them a year or more in the past, the above reliability serves at best as upper bounds for this study.

As far as different perceptions of the responsibilities and resources are concerned, there are examples of obvious inconsistencies. The response number 20 with billet title "Instructor" has allocated 5% to supervising, 5% to representing, 15% to teaching (self-defined item) and 10% to evaluating. Yet only 10% has been allocated to the resource item Personnel. Examples for large profile differences of identical billet titles are numbers 26 and 34, both Flag Lieutenants and Aides, and numbers 83 and 85, Inventory Control Officers and Stock Control Officers, respectively. For the latter example, it has been assumed that both billets are the same. Better agreement of the profiles, especially for the resource items, can be observed for the Detailers, numbers 38, 60, 76 and 95.

In summary, the conclusions concerning the instrument which may be drawn from the direct observations of the data

are as follows. The set of responsibility items seems to embrace nearly all activities Navy Officers encounter in non-operational billets. There are some billets, mainly for junior officers, requiring activities such as teaching, preparation of reports or watch standing. Supplementing the current questionnaire by more specific activities would upset the balance and introduce ambiguity because the items would probably no longer be mutually exclusive, a requirement imposed by the percentage scale. Managerial elements are contained in almost any specific task. Their inclusion would require an arbitrary decision of the respondents where to allocate the time. It follows that the questionnaire with its current set of responsibility items is suitable to classify officer billets of higher echelon.

As far as the resource items are concerned, the responses have indicated that the following changes and additions would improve this set:

1. Change Consumable Supplies to Consumable and Non-consumable Supplies.
2. Add Money and Finances.
3. Add Information and Technical Advice.

Each resource item should also be given a short explanation comparable to those of the other set. In order to make use of the information contained in the set of responsibility importance items and at the same time avoid the weighting problem, a corresponding set should be inserted for the resource items also. A questionnaire originally developed

by the Industrial Relations Center, University of Minnesota, and modified by Professor Boynton, provides for the above proposed improvements. The data should be gathered from officers currently assigned to the billets they describe in order to increase the reliability of the responses.

III. DIMENSIONALITY REDUCTION METHODS

A. PRINCIPAL COMPONENT ANALYSIS

One of the most severe problems in grouping multivariate observations stems from the human inability to visualize configuration of points in more than three dimensions. This is one reason that dimensionality reducing methods are very closely related to classification problems.

Geometrically, every p-dimensional observation vector (p-variables) may be viewed as a point in euclidean p-space. The whole sample is then a cloud of N points. If the coordinate origin is translated such that it coincides with the mean vector of all observations, nothing has changed the configuration of points. Now the principal components of the sample are those new variables which are specified by the coordinates of a rotation of the original coordinate system into an orientation which corresponds to successive orthogonal directions of maximum variance in the data cloud. Analytically, the scores on the principal components are linear combinations of the original variables. They have the property of being uncorrelated with each other and account progressively for smaller variation in the data. Thus, if the first components account for a fairly high percentage (80 to 90%) of the total variance, one may omit the following components by interpreting the last few percent of variation as caused by measurement errors or minor transient effects.

Principal component analysis is a widely used methodology to achieve dimensionality reduction. A more practical reason for applying it is that computer programs which do the cluster analysis are nearly always constrained to process data up to a limited number of objects and variables. Computer time and memory are both highly sensitive to the data volume. Parsimoneous description of data as a scientific principle is always called for.

Principal component analysis can also be used directly to group data in cases when the analysis indicates an essentially low dimensionality (2 or 3) of the data space. Then 1 to 3 scatter plots of the component scores, according to the number of components, can be obtained and the grouping may be performed by visual inspection. This method is the easiest and the most effective classification strategy.

A deeper application of the principal component analysis for the global purpose of grouping is found in the areas of variable selection and factor analysis. Often situations arise where the items of interest for the classification are not directly observable or at least difficult to measure accurately. To circumvent this problem, one could devise indirect, easily measurable items having only the target item in common. Hopefully, the evolving components may be interpreted as those target items by observing high loadings on the corresponding indirect items. It is conceivable that such a procedure may also help to obtain more stable and reliable scores for complex characteristics.

Converting raw scores to component scores may be very advantageous in cases where the variables are highly intercorrelated. As pointed out earlier, correlated variables measure essentially the same characteristic. If no corrections are performed, an implicit higher weight is given to that characteristic when dissimilarities between objects are calculated. The weighting effect of correlated variables can be eliminated by transforming to component scores.

A principal component analysis of the billet data (the responsibility and resource items) has been performed. Its results are shown in Table I.

Table I

Variation Accounted for by the Principal Components

Principal Component	Eigen-Value	% Var.	Cum. % Var.
1	2.47	19.0	19.0
2	2.03	15.6	34.6
3	1.43	11.0	45.5
4	1.24	9.5	55.1
5	1.20	9.3	64.3
6	1.04	8.0	72.4
7	0.90	6.9	79.3
8	0.87	6.7	86.0
9	0.71	5.5	91.5
10	0.57	4.4	95.8
11	0.51	3.9	99.7
12	0.03	0.2	99.9
13	0.01	0.1	100.0

Essentially, the first eleven principal components comprise all information of the data. This result has been expected. It reflects the two constraints imposed by the percentage scale for the two sets of items. Nine principal components are needed to account for 90% of the variance. There is no pronounced decay in the magnitude of the eigenvalues indicating that the swarm of points is shaped fairly spherical. This finding backs the validity of the selected variables and reflects the solid research underlying the emergence of the management functions and the subject areas used in the questionnaire.

As a consequence, it has been decided that the cluster analysis will be conducted on the original variates and not on principal component scores. A reduction from 13 to 8 or 9 principal components was not considered as substantial. In addition, it seems difficult to interpret even very few of the principal components since the original variables themselves have such a high degree of aggregation and concentration.

B. MULTIDIMENSIONAL SCALING

Multidimensional scaling is a technique which is widely applied in the fields of psychology and sociology. It constructs a configuration of points in p-space from information about the mutual dissimilarities of the points. The target dimensionality p has to be specified. The dissimilarities need not qualify as metrics, satisfying the triangle inequality.

They may be derived from measurements or obtained directly, for example, by human comparative judgment. In short, multidimensional scaling can recover metric information from non-metric inputs.

The resulting configuration of points is constructed such that a monotonic relationship between the original dissimilarities and the corresponding interpoint distances is satisfied as closely as possible. Kruskal in his important paper [8] credited Shepard to be the one who showed that the rank order of the dissimilarities is itself enough to determine the solution. This is true because a matrix of dissimilarities imposes so many constraints on the locus of points that the resulting configuration is completely determined by knowing only the rank order of the matrix entries.

The algorithm accomplishing the goal of recovering the original configuration works as follows. Let us assume an nxn dissimilarity matrix has been obtained for n objects. The algorithm starts with an arbitrary configuration in the prespecified p-space. A criterion called "stress" is then computed. It measures how well the rank ordering of the interpoint distances fits to the rank ordering of the dissimilarities and may be interpreted as a badness of fit measure. In an iterative fashion using the method of steepest descent from non-linear programming, all points are moved such that the stress becomes smaller. The algorithm stops when no further movement can decrease the criterion.

It is intuitive that the stress of a higher dimensional representation is always less than that of a lower dimensional representation because in higher dimensions there are more degrees of freedom for moving the points. If the number of dimensions is greater than or equal to n-1, the perfect stress of 0 can always be achieved.

To decide on the appropriate number of dimensions, it is common to obtain several configurations in different dimensional spaces and plot the minimum stress versus the number of dimensions. A noticeable elbow in the curve where a further increase in the dimensionality results only in a small decrease of the stress, gives an indication of the effective number of dimensions. In general, the bigger the size of the dissimilarity matrix the more dimensions may be extracted. Another criterion which may influence the number of dimensions is the interpretability of the axis in cases where such an interpretation is essential for the analysis. When it may be assumed that the error portion in the dissimilarities is large, fewer dimensions are called for.

For the classification problem, multidimensional scaling can be applied in two rather different ways. It may serve as a dimensionality reduction method which in contrast to principal component analysis is not based on a linear model. The second application is the determination of the basic underlying dimensions in judging billet similarities (or dissimilarities). Hypothetically, officers concerned with personnel management could be asked to judge directly from

a set of billets the similarity of any possible pair on an ordinal scale. Multidimensional scaling then could generate a configuration in p-space. A rotation of the axes may be performed such that the coordinates can be interpreted meaningfully as the basic underlying judgment dimensions. This kind of application has not been used in this study. It has only been mentioned to give a more complete picture of this interesting methodology.

The billet data has been submitted to a multidimensional scaling program solely for exploratory reasons. The classification problem did not require the recovery of metric information because the scale of the items allows the direct computation of any metric. In addition, the 13 dimensions of the raw data are not intractable by numerical classification methods. Yet, a comparison with the results of the principal component analysis could give further insight into a possibly lower dimensional representation.

A dissimilarity matrix $[\delta_{ij}]$ has been computed from the raw scores by using euclidean distance

$$\cdot \quad [\delta_{ij}] = ||x_i - x_j||, \quad i,j = 1,\ldots,93$$

where

$$x_i = (1 \times 13) \quad \text{data vector of billet } i.$$

The scaling has been performed by a computer program, called KYST. KYST incorporates ideas from Kruskal, Young, Shepard and Torgensen whose initials form the name. An obstacle had to be overcome in order to apply the program. The lower half

of a 93 x 93 matrix without the diagonal has 4,278 entries. The capacity of the program arrays allows for a maximum of 1800 data values. A cutoff point for the distance values had to be chosen such that all entries below that value are disregarded. Finally, 1761 values have been read in comprising 41% of the total number. Kruskal has pointed out that for large (greater than 13 x 13) matrices the loss of information is not significant even if 75% of the dissimilarities are omitted. Configurations in 2 to 6 dimensions have been obtained. The plot of stress versus the specified number of dimensions is shown in Figure 1. It can be seen that the plot exhibits an elbow at 5 dimensions, indicating that the data may be reduced to 5 dimensions. The achieved stress of .028 associated with that configuration may be ranked as excellent due to Ref. [8]. This finding is in partial accord with the result of the principal component analysis where it is a common heuristic to retain only components which are associated with eigenvalues greater than one. Its application would call for a 6-dimensional representation, and 75% of the variance would be accounted for.

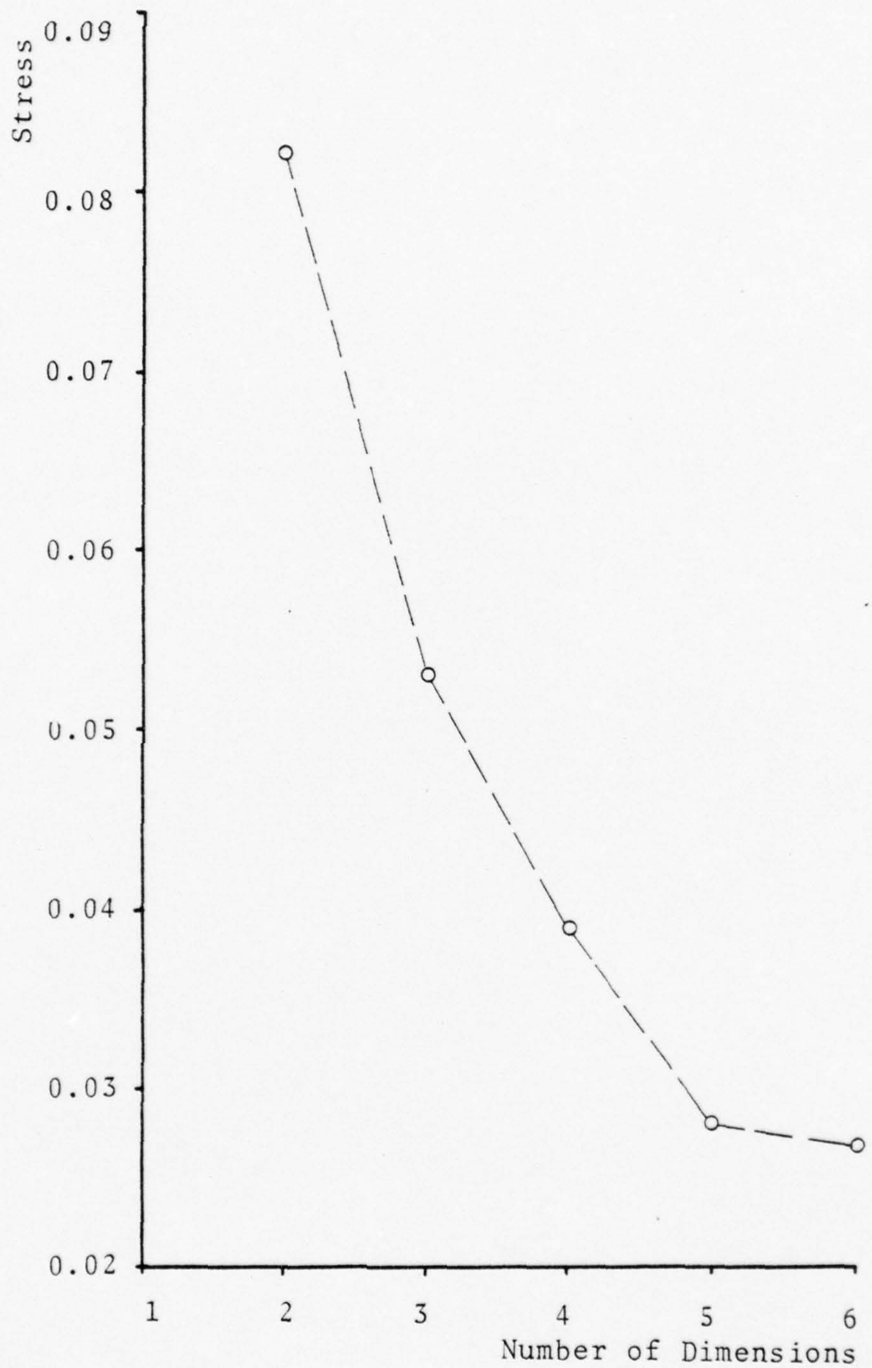


Figure 1. Plot of Stress vs. Number of Dimensions

IV. CLUSTER ANALYSIS

A. CONCEPT

Unlike such techniques as discriminant analysis and statistical decision theory where the investigator starts from an external model, cluster analysis does not provide for such an external reference. The only belief is that there should be meaningful subgroups in the sample of N objects. However, because of the lack of an external criterion with which to define these subgroups, the researcher tentatively adopts an internal criterion by letting the data itself determine "natural" groups. As such, the primary value of cluster analysis lies in the preclassification of data.

Any method which partitions a set of objects into subsets on the basis of measurements taken on every object will be called a clustering process in this study. The partition into subsets, called clusters, is done in such a way that every object belongs to a cluster and that the objects in a particular cluster are most similar with respect to the variables measured.

Cluster Analysis has become a prominent tool for analyzing multivariate observations. Some of its goals are:

- data reduction
- data exploration
- hypothesis generating
- finding a true typology.

Most of the early applications of clustering techniques have been in the fields of biology and zoology under the name of Taxonomy. It is important to obtain a clear intuitive picture of the concept of cluster analysis in order to appreciate when it may be of potential value in specific instances. The geometric model which has been introduced earlier is very well suited to serve this purpose. Due to that model, every object can be viewed as a point in a p-dimensional euclidean space. Now it is conceivable that the whole swarm of data points contains some continuous dense regions or "clouds" which are separable from other regions containing a relatively low density of points. This intuitive definition of clusters does not restrict the shape in any way.

In two dimensions it is easy for the human eye to identify denser regions from a scatter plot. When more than two or three dimensions are involved, the finding of clusters is much more difficult and nearly always performed with the help of electronic computers.

With the uprise of computers in the last 20 years, the number of algorithms capable to detect cluster has grown at a very fast rate. Yet so far, clustering techniques are lacking the desirable statistical and mathematical rigor. As a consequence, the researcher has to make several arbitrary decisions in the course of an analysis which will influence the cluster solution. The most important decisions are:

1. Which algorithm to apply?
2. What to use as a measure of dissimilarity?
3. How many clusters are there?

4. Should the data be normalized?
5. Should the variables be weighted differently?

Most of the above questions are intimately related and may not be answered satisfactorily in general, but only in context with the problem at hand. For this study, the choices will be delayed till the next two sections of this chapter.

It is clear that every clustering procedure results in a partition. One of the most difficult problems associated with an analysis involving a clustering process is to determine whether the resulting clusters are significant or whether the partition is the result of imposing artificial structure onto the data. Up to now, there appears to be no robust statistical test available which might answer this important question.

The choice of algorithm for the present study had to be made between a hierarchical clustering method and a scheme which optimizes an objective function. The hierarchical cluster program has been developed by Johnson [7] and is called "Hiclust"; the other one called "K-Means Iterative Clustering Program", has been developed by McRae [11].

B. MEASURES OF DISSIMILARITY

A measure of similarity or dissimilarity between objects is needed in order to classify objects into groups. In general, this is a function defined on any pair of objects which maps the pair into a real positive number. If x is the nxp data matrix, where n is the number of objects and p is the number of measurements (variables) taken on each object, then

$$[\delta_{ij}] = f(x_i,x_j), \quad i,j = 1,\dots n$$

where

f = dissimilarity function

$[\delta_{ij}]$ = nxn dissimilarity matrix

x_j = (1xp) data vector of object j.

The entry δ_{ij} is small when x_i and x_j are similar.

The most important dissimilarity functions are metrics. They satisfy the following 4 properties:

1) $d(x_i,x_i) = 0$

2) $d(x_i,x_j) \geq 0$ for all x_i, x_j

3) $d(x_i,x_j) = d(x_j,x_i)$

4) $d(x_i,x_k) \leq d(x_i,x_j) + d(x_j,x_k)$

Condition 4 is, of course, the well-known triangle inequality. Although the geometric model of points in p-space implicitly assumes the dissimilarity function to be the euclidean distance measure, f has not to be a metric in general.

Two other groups of dissimilarity functions are correlation coefficients and agreement coefficients. Sokal and Sneath [12] have described several of the latter coefficients and their properties. It is essential that the choice of the dissimilarity function is compatible with the scale of the variables. Gover [4] has presented a general coefficient of similarity which is defined for any scale. He has shown that a matrix of that particular coefficient is positive semi-definite.

In the following, some of the most widely used dissimilarity measures for quantitative variables are listed and defined.

Euclidean distance:

$$d(x_i, x_j) = \|x_i - x_j\|$$

Mahalanobis distance:

$$D^2(x_i, x_j) = (x_i - x_j)T^{-1}(x_i - x_j)'$$

T is the total scatter matrix or matrix of cross products defined as:

$$T = \sum_{i=1}^{n} (x_i - m)'(x_i - m)$$

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Dissimilarity measures based on the product moment correlation:

$$r_{ij} = \frac{x_i \cdot x_j}{(\|x_i\|^2 \cdot \|x_j\|^2)^{\frac{1}{2}}}$$

Written in this form, it is assumed that the object vectors have been centered so that the elements add to 0.

L_1-norm (City Block Metric):

$$L_1(x_i, x_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

x_{ik} = k'th element of observation i.

Euclidean distance is the most intuitive metric. It leads immediately to the model of n points in p-space. One can easily visualize its invariance under a rotation of axis.

Whenever the cluster-conditions are ideal in the sense of nearly uncorrelated variables, equal scale of measurements and balanced variation of the variables, this measure is the natural one to apply.

Mahalanobis distance, also known as statistical distance, has very appealing properties. It can be shown that this metric is not only invariant under an orthogonal linear transformation (rotation of axis), but invariant under any linear transformation [2]. It was mentioned earlier that the linear transformation of the data (corrected for means) to principal components eliminates the implicit weighting effect of correlated variables. Applying the statistical distance achieves the same goal. The total scatter matrix T, whose inverse serves as a weighting scheme, is proportional to the well-known sample covariance matrix.

Despite its invariance property, the Mahalanobis distance has a serious drawback if it is applied to clustering problems. It tends to decrease the clarity of existing clusters. Hartigan [5] discusses this important point more thoroughly. The same tendency, but with a less degree, exhibits euclidean distance where every term has been scaled by the inverse of the corresponding sample variance,

$$d_s(x_i,x_j) = \left(\sum_{k=1}^{p} \frac{1}{S_k{}^2} (x_{ik} - x_{jk})^2\right)^{\frac{1}{2}}$$

where S_k = sample standard deviation of variable k.

Euclidean distance scaled in this way is the same as ordinary euclidean distance applied to data which has been standardized to have equal variance. By this procedure, the data becomes invariant to the units of measurements and all variables contribute equally to the squared distance.

The blurring effect of scaled distance measures can be circumvented by using an average of the within-cluster variances as weighting factors. Unless estimates of the within-cluster variances obtained from an independent source are available, a circular reasoning becomes apparent: A distance measure is needed to identify groups in the data. Knowledge of the group membership is required to obtain a useful distance function.

The problem has its exact analogy for the Mahalanobis distance, if the inverse of the total scatter matrix is replaced by the inverse of the within-scatter matrix W, defined as

$$W = \sum_{j=1}^{g} W_j$$

where W_j is the cross product matrix of cluster j.

Both "non-diluting" forms of weighted distance cannot be applied in the hierarchical clustering process because they require knowledge of the grouping one is looking for.

For any other weighting scheme the following rule applies. All items are to be equally weighted when there is no a priori information of their relative importance. Once a classification has been obtained and the evolving clusters are further

investigated, it might be possible that as a result of the preclassification some variables turn out to be more important than others and should therefore have assigned greater weights.

Dissimilarity measures based on the product moment correlation between profiles may be advantageous in special cases. Geometrically, the product moment correlation is the cosine of the angle between two points (vectors) assuming the elements have been centered by subtracting its mean. A small, angular separation between those points gives a small dissimilarity measure, no matter how distant the points are located from the origin. Thus, only the shape of the profiles and not the magnitude determines their mutual dissimilarities.

The City Block Metric is appealing when the scale of variables is strictly a percentage scale. The data profiles are constrained by this scale to sum up to a constant. To see the effect of euclidean distance applied to data on a percentage scale, consider the two pairs of extreme profiles in eight variables:

a)	25	0	25	0	25	0	25	0
b)	0	25	0	25	0	25	0	25
c)	100	0	0	0	0	0	0	0
d)	0	100	0	0	0	0	0	0

It may be argued that (a) is as dissimilar from (b) as (c) is from (d). Yet, euclidean distance gives

$$d(a,b) = 50\sqrt{2} \text{ and}$$

$$d(c,d) = 100\sqrt{2},$$

which seems not at all to reflect the "true" relations. The more numerous the variables, the bigger will be the difference between both extreme pairs. City block metric results in a distance of 200 for both pairs. In the case of an uneven number of variables, a similar effect can easily be verified. One might generalize that a natural metric for data on a percentage scale is the city block. It is a true metric, but it is not invariant under a rotation of the coordinate axes. Also, it is not naturally coupled with an inner product.

The discussion about the various dissimilarity measures may be summarized as follows: The choice is highly dependent on the scale of the data. When there is no danger that some variables overpower others, and the geometric model of points in p-space seems valid, euclidean distance based on the raw scores should be used. Blindly standardizing the data prior to computing distances without a critical judgment of its need may harm the clarity of possibly significant clusters. For the same reason, Mahalanobis distance based on the total scatter matrix is not necessarily the best choice of a distance function.

C. HIERARCHICAL CLUSTER ANALYSIS (HICLUST)

Hierarchical clustering methods are probably the most widely used of all clustering algorithms. They are very efficient in terms of computer time. The first step in a hierarchical cluster analysis is to convert the data matrix into a dissimilarity matrix by applying a dissimilarity function.

The next step in a hierarchical cluster procedure involves manipulation on the dissimilarity matrix such that the partition into subsets may be visualized easily. The vehicle for visualizing clusters is the dendrogram. An example of a dendrogram for five objects is given in Figure 2.

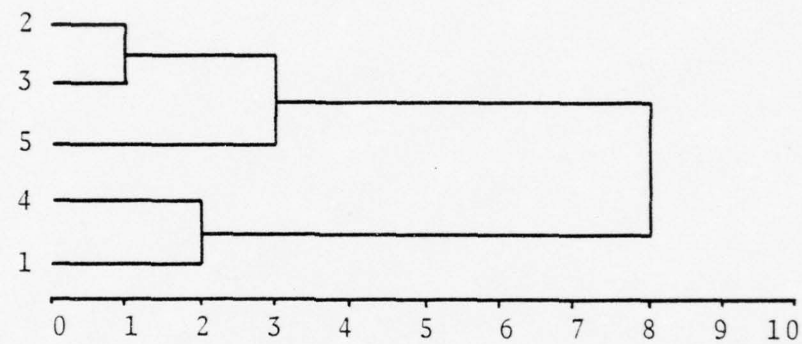


Figure 2. Example of a Dendrogram

At level 0 every object constitutes a cluster. The first merging happens at level 1 where objects 2 and 3 form a cluster. With rising level, more and more objects merge until finally at level 8 all objects are fused together. The dendrogram in Figure 2 probably suggests a two-cluster solution with one cluster comprising the objects 1 and 4 and the second cluster consisting of objects 2, 3 and 5. A large jump of the merging level is an intuitive heuristic where to draw the dividing line between clusters.

There are several possible ways to construct a dendrogram from a matrix of dissimilarities [9]. Two particularly useful

algorithms are known under the names Single Link and Complete Link which have been implemented into Hiclust. Both belong to the class of agglomeration algorithms. The dendrogram is constructed by a series of successive fusions of smaller clusters into larger clusters, starting with each object being a cluster until all objects are finally grouped together. The Single Link procedure generates clusters which tend to be long and stringy. The complete link algorithm produces compact clusters which are relatively dense with respect to the surrounding clustering space. In Ref. [7] it is shown that both methods are invariant under any monotonic transformation of the dissimilarity measure. Furthermore, it is shown that there is a one-to-one correspondence between a dendrogram and an ultrametric. Ultrametrics satisfy the ultrametric inequality

$$d(x_i,x_j) \leq \max\{d(x_i,x_k),\ d(x_j,x_k)\}$$

which is a stronger condition than the triangle inequality. Because of the one-to-one correspondence between a dendrogram and an ultrametric, every hierarchical cluster algorithm may be viewed as a procedure which transforms a dissimilarity matrix (not necessarily based on a metric) into a matrix of ultrametrics.

Each of the above described steps towards representing the hierarchical structure of the data involves a distortion. The choice of the dissimilarity measure and the choice of the algorithm to transform dissimilarities to an ultrametric

influence the degree of the distortion. Of course, the more pronounced clusters are contained in the data, the greater is the likelihood to "survive" and to re-emerge finally in the visual representation.

Hierarchical methods are especially useful for data with variables measured on vastly different scales. If, for example, some variables are measured on a ratio scale like relative amount of time spent in fulfilling a certain responsibility, and other variables merely indicate the absence or presence of a particular characteristic as is often the case in task inventories, hierarchical clustering methods are easy to apply. The reason is the two-step process of forming a dissimilarity matrix first and subsequently transforming it to an ultrametric. Another feature of the two-step process is that large amounts of information collected on each object may be summarized by reducing it to a single dissimilarity index. Thus, there is no need for a prior dimensionality reduction of the data. On the other hand, the application of hierarchical scheme is constrained by the number of objects which can be classified. The storage requirements increase quadratically with that number. Hiclust is limited to process a 100 by 100 matrix.

A test application of Hiclust to data with known structure and a comparison of the results with those obtained by K-Means has been included as Appendix C. As a result of this test application, the conclusion has been drawn that for very well structured data, in the sense of the presence of distinct

groups, Hiclust is a fast and efficient data exploration tool. When this is not the case, Hiclust is likely to give misleading results.

In applying Hiclust to the billet data, it has been decided to use the Complete Link algorithm in order to avoid the chaining effect of the Single Link algorithm. A dissimilarity measure based on the product moment correlation has been chosen. The reason is that some profiles do not add up to 200% because of time allocation to self-defined items. The selected dissimilarity has the advantage to be least affected by that inconsistency. Figure 3 displays the obtained dendrogram.

As Hiclust indexes the objects by their respective position from the dissimilarity matrix, the previously discarded sample units 22, 47 and 58 have been included to maintain the one-to-one correspondence between index and billet.

From the dendrogram it can be seen that at a level of .70, ten clusters have formed. The dividing lines from top to bottom are between billets 78/69, 84/43, 79/63, 95/16, 58/13, 75/26, 93/33, 94/22 and 28/41. The ten clusters obtained by Hiclust are not further displayed and discussed because there are several indications that the billet data contains mainly unpronounced subgroups. Under these assumptions, it has been seen that Hiclust tends to create artificial clusters.

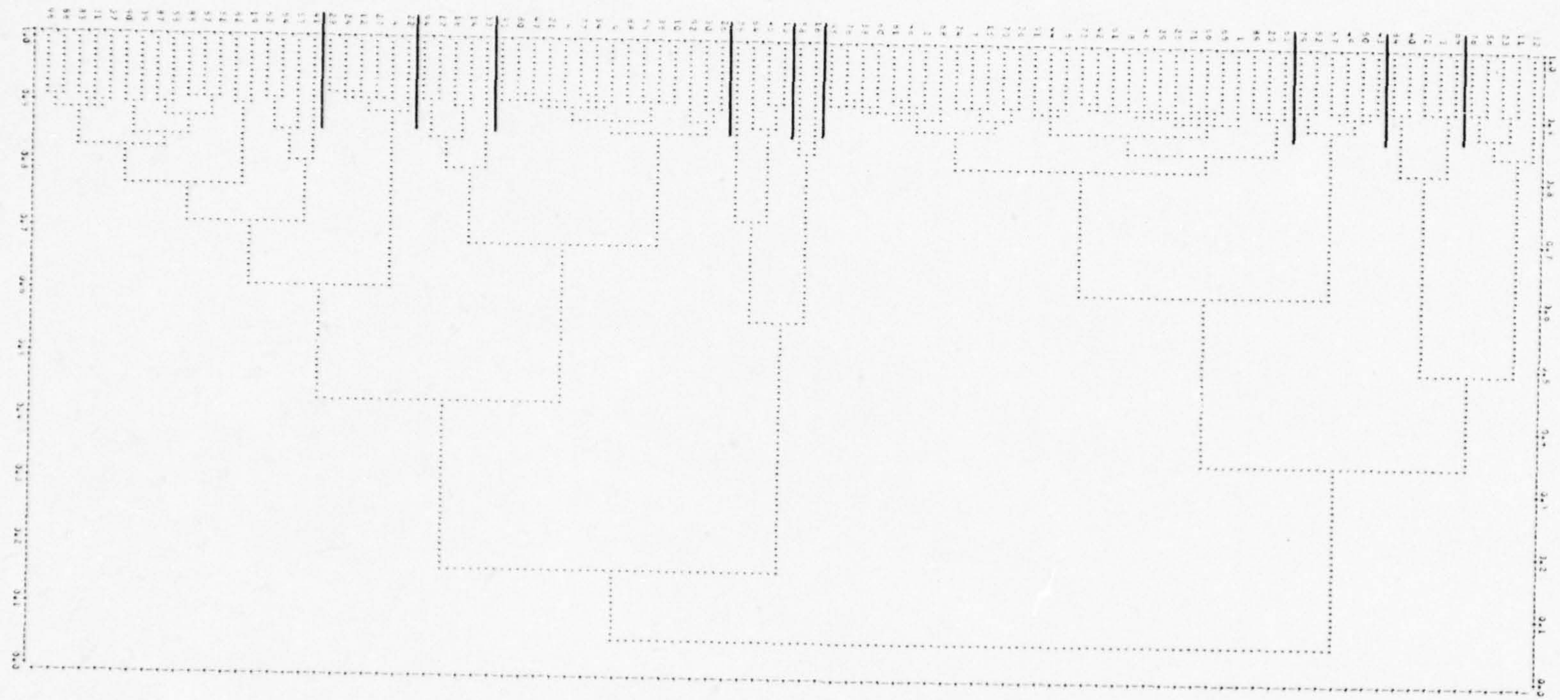


Figure 3. Dendrogram of Billet Data

D. CLUSTER ANALYSIS OPTIMIZING AN OBJECTIVE FUNCTION (K-MEANS)

1. General Features

The main feature of the K-means algorithm is that it is seeking a grouping of the data which optimizes a real valued function, defined on any partition. The basic iterative procedures work as follows. The algorithm starts with g arbitrary points (g is the specified number of clusters), serving as initial cluster centroids. Then, in the order of input, each point is assigned to that cluster centroid which is nearest and the centroid is recomputed. After one pass through of the data the criterion value of the resulting partition is calculated and the process starts again with the assignment of the first point unless the criterion value has not improved. In that case, the second iterative step begins. Each point, starting in cluster 1, is reconsidered in a specific order within each cluster to be moved to another cluster with the move being made only if the criterion value improves. When no move of a point can improve the objective function, the algorithm stops.

Quite different from Hiclust, K-Means does not manipulate the matrix of dissimilarities. Instead, it works with the original observation vectors themselves. Thus, at any stage of the iteration, the ability to incorporate changes of weights according to the variance-covariance structure of the current cluster configuration is maintained. This feature enables the algorithm to circumvent the above described

circularity in the use of weights which take into account the within-cluster variance or within-scatter matrix. Thus, scaled euclidean distance and Mahalanobis distance may be specified without encountering the diluting effects of using the estimates of total variance and the total scatter matrix, respectively, as weights.

K-Means offers the following options.

Distance functions:

- Euclidean distance
- Scaled euclidean distance
- Mahalanobis distance

Objective functions:

- Trace (W)
- Determinant (W)
- Largest eigenvalue of $W^{-1}B$
- Trace $(W^{-1}B)$

The distance functions and some of its most important properties have been described in the last section. Next, the criterion functions will be illuminated shortly. Friedmann and Rubin [3] discuss this topic more thoroughly.

The objective functions have their origin in multivariate analysis and are all based on scatter matrices. The well-known partition of the total sum of squared error can be generalized to the matrix identity

$$T = W + B$$

The total scatter matrix T and the within-scatter matrix W have been defined earlier. The between-scatter matrix B is defined as

$$B = \sum_{j=1}^{g} n_j (m_j - m)'(m_j - m),$$

where

m = $(1 \times p)$ centroid of the total sample

n_j = number of objects in cluster j

m_j = $(1 \times p)$ centroid of cluster j.

The trace (W) criterion is more commonly known as the Sum-of-Squared-Error criterion J_e and is defined as

$$J_e = \sum_{j=1}^{g} \sum_{x \varepsilon X_j} ||x - m_j||^2 .$$

X_j = Set of objects in cluster j, j=1,...,g.

It is the sum of squared distances of every point in a cluster to its centroid pooled over all clusters. Ideally, every point in a homogeneous cluster should be representable by its centroid. Therefore, the distance of a point to its cluster centroid represents the deviation from homogeneity. Minimizing this criterion will tend to produce spherically-shaped clusters with about equal variation in all variables.

Another less obvious feature of the trace (W) criterion becomes apparent when there are big differences in the number of points in adjacent clusters. Splitting the more populated but quite dense cluster often results in a greater decrease of the criterion than does the true partition into a large and a small cluster. Scaling of variables will change the partition obtained by optimizing this function. Also important to note is its implicit assumption of the euclidean distance measure. Partitions are therefore invariant under a rotation of axes.

The other three objective functions have their origin in multivariate analysis and are related to the eigenvalues of

$$W^{-1} B.$$

In order to find compact clusters, Det (W) has to be minimized and the other two functions of the partitions have to be maximized. In contrast to the Sum-of-Squared-Error Criterion, the multivariate criteria have in common the property of invariance under any linear transformation. In addition, these criteria are less restrictive concerning the assumption of spherically-shaped clusters. They allow for different hyper-elliptical shapes of the clusters [3].

The test application of K-Means to the data with known structure has strongly indicated that optimizing one of the multivariate objective functions will lead to better solutions than by optimizing the trace (W) criterion. It revealed also the negative effect of prior standardizing the data. Investigations about the influence of shortening the iterative cycle have further shown that the Mahalanobis distance gives the fastest rate of convergence [11].

Multivariate criteria are applicable only if the within-scatter matrix W is non-singular. Assuming that no linear relation exists between the p variables, W is non-singular as long as $p \leq n - g$ where n is the sample size and g the number of clusters [13].

The percentage scale imposes a linear constraint upon the scores. Centering any subset of objects by subtracting its centroid will result in a linear dependent set of variables. Consequently, the matrices W_j have a rank of at most p-1. This may lead to a singular matrix W, a consideration that needs attention in future development of the method. This point posed no problem in the current study because the self-defined components were omitted from both the responsibility items and the resource items.

The singularity of W can be avoided by a linear transformation of the data to principal components. For the case of linear constraints, the space of components is reduced by the number of constraints. No loss of information is incurred because merely redundancy is removed.

The use of K-Means to obtain an hierarchical cluster structure has proven to be limited as far as the billet data is concerned. There is no forcing into hierarchical structures comparable to Hiclust. Generally, when the cluster number is increased from g to g+1, more than one of the g-clusters contributes to the formation of the additional cluster by releasing some of its members. Yet, the application of K-Means to the test data has displayed the inherent hierarchical structure of that data very well. It has been observed that in the presence of well-structured clusters, K-Means also has the capability to indicate the hierarchical relations.

2. Application of K-Means to the Billet Data

For the cluster analysis of the billet data, it has been decided to use the Mahalanobis distance measure and to minimize the determinant of W. This criterion has proven to be as powerful as both other multivariate criteria. Computationally, it is faster. As the units of measurements are the same for all items, it has been decided to perform the analysis on the raw scores instead of prior standardizing.

Next, the number of clusters had to be specified as an input parameter for K-Means. The lack of an outside model or any other preconception about the number of groups present called for a decision rule which had to be generated by the data itself. An intuitive heuristic is to perform the analysis for several different cluster numbers and then to plot the achieved optimal criterion value versus the number of clusters. If there is a number such that the marginal improvement of the criterion becomes insignificant from then on, this number would be the appropriate choice.

In some computer runs it has been observed that by specifying a larger number of clusters, the criterion value even increased again. The reason may be due either to computational noise or the lack of convexity of the criterion function. Consequently, there is no guarantee that the K-Means algorithm converges to that partition which results in a global minimum value for the function. The optimal cluster solution depends on the random selection of the starting centroids. Therefore, several computer runs have

been performed with a randomized order of input to increase the probability of convergence to a global solution.

In Figure 4 the logarithm of the determinant of W has been plotted versus the number of clusters. Up to eight clusters, the optimal value decreases fast. From then on, the criterion stays nearly constant, suggesting the presence of eight groups.

The associated partition has been selected as the final solution and is included as Appendix D. The group centroids and the within-group standard deviations for each cluster are given. The centroids are plotted.

Figure 4. Plot of the Objective Function Log Det(W) vs. the Number of Clusters

V. SUMMARY

A. METHODOLOGIES

Principal Component Analysis and Multidimensional Scaling serve mainly to reduce the dimensionality. Such is often a useful preclassification step. The latter method has the additional capability of constructing a low dimensional spatial configuration even when the data is non-metric originally. In certain fortunate situations they may result directly in a classification of the objects. This is the case where two- or three-dimensional representations may be obtained which account for most of the originally present information. So far no algorithm can excel the human brain for grouping objects from a scatter plot in two dimensions. Modern graphical computer output allows the visualization of scatter plots in three dimensions. Whenever low dimensional representations are possible, the application of numerical classification methods should not be considered.

If no such low dimensional representation is possible, then clustering schemes become more valuable. The hierarchical procedure is advantageous when well separated groups of objects may be assumed to exist in the data. The geometric model of points in euclidean space is not needed. The method is most powerful if the number of measurements (variables) taken on the objects is large. Prior dimensionality reduction

steps are not required; moreover, they are not even desirable in connection with this method.

One of the most appealing properties of hierarchical cluster analysis is its robustness with respect to the scale type of the measurements. Data of any scale, from nominal to ratio, can be processed. The use of weights and even the presence of hierarchical relationships between different variables present no problem in applying this kind of analysis. The graphical output obtainable from this method may give valuable information about the hierarchical structure of the groups.

The drawback of Hierarchical Cluster Analysis is its relative crudeness. The steps towards generating the conspicuous output distort the information content of the data. Its application is limited by the number of objects to be classified.

K-Means is the more sophisticated classification method optimizing an objective function. It applies to data consisting of points in euclidean space. Thus, if the scales of measurements do not meet the requirements of the geometric model, the method cannot be applied directly to the raw data. Multidimensional scaling may serve then to transform nonmetric input into metric information. This process, however, imposes also some distortion to the data.

The application of K-Means is most advantageous when the number of objects to classify is large and the dimensionality is small. Furthermore, this method gives a valuable heuristic

indication of how many groups are represented in the sample. An application of K-Means to a data set of known structure has proven its capability to separate groups which Hiclust did not identify. The gain of information about the hierarchical relations among the groups, however, is limited.

B. CLUSTER SOLUTION

The groupings obtained vary in size from 3 (cluster 5) to 29 (cluster 2) and exhibit quite different stability behavior. Clusters 1, 2, and in a less degree, cluster 3 remained nearly unchanged as the number of clusters extracted by the K-Means program advanced from 6 to 10, indicating a higher level of homogeneity. The first two of these "stable" clusters can also be recognized from the dendrogram in Fig. 3. Clusters 4 to 8 seem to be less dense. Their composition changed drastically with increasing cluster number.

An investigation of the cluster composition revealed that billet titles appear to be of little use for classification purposes when managerial responsibility and resource items are the input characteristics of the classification analysis. There seems to be greater homogeneity of responsibilities existing within the identified classes than within any common grouping by billet title. Even billets with the same title, but from different commands and locations, may exhibit rather different profile scores. It can be seen from the within cluster standard deviations in Table II that the items discriminate quite differently for the various groupings.

The sums of squares of the within-cluster standard deviations over all items yield the smallest values for clusters 1 and 2. This is another sign that they are relatively more compact.

Table III gives an indication of the overall discriminating ability of the items for the obtained cluster solution. The resource items seem to discriminate better than the responsibility items.

The next developmental step should apply these methods of analysis to a more representative sample which should be large so that the cross-validation technique can be applied. The current instrument appears to be competently designed. Its main area of application should be limited to high ranking officers because junior officers often have to fulfill specific tasks. These tasks are difficult to describe accurately by the responsibility items of the questionnaire. This is believed to be the main reason that the current data is lacking the desirable structure in the sense of well pronounced subgroups.

To achieve the long range goal of developing a more detailed billet classification for a specific purpose, the input information has to be closely related to that purpose. This requires that future data collecting tools should have high resolutive power by providing for specific purpose-related items.

APPENDIX A: QUESTIONNAIRE

Professor R. S. Elster, CodeEa

Professor R. R. Read, Code 55Re

Officer Responsibilities Questionnaire

BACKGROUND

1. We are working on a project addressing the tasks performed by Naval officers in their jobs. The research is sponsored by the NPS Foundation Research Council and is funded by ONR.

2. This questionnaire is our first broad-brush effort at garnering job description information from officers. In later iterations of this sort of questionnaire, we will ask other officers for more information in job responsibility areas that emerge as important from your responses today.

3. Eventually, we hope to develop a method for gathering data describing officers' billets that will help the Navy to determine which billets should be P-coded, groups of billets which are very similar to another, and so on.

4. Thank you for your assistance.

GENERAL INSTRUCTIONS

1. We would like you to describe the last non-student, non-operational billet you had prior to coming to NPS. (If you have not had such a billet, just hand this back without completing it.)

2. The attached questionnaire includes questions about that billet and the responsibilities you had in that billet.

Please write the last four digits of your Social Security number here _______

Title of the billet you are describing:

1. On the next page are definitions of eight management responsibilities. Please read over all the eight responsibility descriptions, and then:

   a. Considering the entire range of activities over your tour of duty in that billet, estimate the average percentage of your time you spent on that area of responsibility. Write that percentage on the line to the right of the responsibility. Please write the percentages as follows:

      - write 5 as 005
      - write 50 as 050
        etc.

      - No decimals, please.

   b. The final category, "Other", is provided in case you had other responsibilities in your billet.

   c. The percentages you write down should total to 100.

   d. Again, please scan over all eight responsibility areas before you begin to write down any percentages.

   e. You probably should use a pencil so you can make easy changes to your answers.

Key punch
Columns

Please write the last four digits of your Social Security Number ________ 1-4

Your rank when you were in that billet. Write in one number on the line to the right. Let: 1 = 0-1, 2 = 0-2, 3 = 0-3, etc.; write a 1 if you were an O-1, etc. ______ 5

Responsibilities

1. Planning: Determining goals, policies, and course of action to be taken.

   - Examples of tasks and products included in Planning: Training plans, work scheduling, budgeting, setting up procedures, setting goals and standards, preparing schedules, preparing op. orders, career planning, etc. ______% 6-8

2. Investigating: Collecting and preparing information for reports (oral or written), records, or accounts.

   - Examples of tasks and products included under Investigation: inventorying, financial records, legal, job analysis, record keeping, doing research, incident investigation, screening, trouble-shooting, etc. ______% 9-11

3. Coordinating: Exchanging information with people in the organization other than your subordinates in order to expedite, adjust work problems, and ensure proper completion of assignments.

   - Examples of tasks and products included under Coordinating: advising other departments, units, commands; liaison with other officers or civilian managers; expediting the accomplishment of something; arranging meetings; informing or briefing peers or superiors; performing inter-service liaison, etc. ______% 12-14

Key punch Columns

4. Supervising: Directing, leading, and developing your subordinates.

   - Examples of tasks and products included under Supervising: assigning work, training subordinates, handling complaints, directing work/watch, counseling subordinates, disciplining, etc. ______ % 15-17

5. Evaluating: Assessment and appraisal of personnel, equipment, and proposals.

   - Examples of tasks and products included under Evaluating: personnel appraisals, personnel inspections, performance appraisals, approving requests, approving plans, reports, records; inspecting equipment, other inspections, judging proposals or suggestions, serving on pilot disposition boards, evaluating intelligence data, etc. ______ % 18-20

6. Staffing: Maintaining the proper numbers and kinds of personnel in your department, unit, command, of several units, etc.

   - Examples of tasks and products included under Staffing: securing needed personnel; assigning, promoting, transferring personnel; personnel retention efforts; reducing required number of personnel; preventing personnel piracy, etc. ______ % 21-23

7. Negotiating: Either formal negotiations outside the service in contracting for goods or services, or "maneuvering" within the service or government to obtain resources.

   - Examples of tasks and products included under Negotiating: dealing with sales representatives; contacting suppliers either within or outside of the service; collective bargaining; "making noises" to obtain or expedite assignment of personnel, funds, supplies, or services; negotiating trade-offs of resources between departments, units, etc. ______ % 24-26

Key punch
Columns

8. Representing: Advancing the general interests of your organization through speeches, consultations, and other activities or contacts with individuals or groups outside of your unit, department, or command.

   • Examples of tasks and products included under Representing: making public speeches, community activities, international social activities, presentations, briefings, conducting tours.

   ______ % 27-29

9. Other management responsibilities. (Please specify what they were.)

   ______ % 30-32

Please check to make sure the eight or nine percentages you wrote down above total to 100%.

We would like to obtain your views concerning the relative importance of the responsibilities in the job you are describing. In other words, we wonder if performance in some of the responsibility areas might be more, or less, important than the time percentages you gave above would indicate.

We would like you to allocate 100 points among the eight or nine responsibility areas. The responsibility given the greatest number of points should be the one you feel was most important to success in your job. Allocate the 100 points in such a way that the ratio of the points allocated to responsibilities matches your perception of their relative contributions to success in the billet. If, for example, you feel Planning was twice as important as Investigating in determining success in your job, allocate twice as many points to Planning as you do to Investigating.

Please write the points for each responsibility on the line to its right. Please write the points as follows:

- write 5 as 005
- write 50 as 050

  etc.

- No decimals, please.

Responsibility	Points	Key punch Column
Planning	______	33-35
Investigating	______	36-38
Coordination	______	39-41
Supervising	______	42-44
Evaluating	______	45-47
Staffing	______	48-50
Negotiating	______	51-53
Representing	______	54-56
Other (as previously specified)	______	57-59

Total Points must sum to 100. Total: ______

In the billet you are describing you exercised your responsibilities over resources such as people, equipment, facilities, consumable supplies, and over methods and procedures, etc. Please estimate the percentage of your time you spent exercising responsibility over each area. Your estimates should total to 100%. Please use only two digits for each percentage. For example:

- write 5 as 05
- write 50 as 50

  etc.

- No decimals, please.

Resources and Subject Matter	Time Spent	Key punch Columns
Personnel	________	60-61
Equipment	________	62-63
Facilities	________	64-65
Consumable Supplies	________	66-67
Methods and Procedures	________	68-69
Other (please specify)	________	70-71
	________	72-73
	________	74-75
Total should be 100. Total=	________	

APPENDIX B: DATA BASE

1. Rank Indexes

0	Warrant Officer
1	Ensign
2	Lieutenant J.G.
3	Lieutenant
4	Lieutenant Commander
5	Commander

2. Responsibility (R) and Responsibility Importance (RI) Item Indexes

R	RI	
1	10	Planning
2	11	Investigating
3	12	Coordinating
4	13	Supervising
5	14	Evaluating
6	15	Staffing
7	16	Negotiating
8	17	Representing
9	18	Other

3. Resource Item Indexes

19	Personnel
20	Equipment
21	Facilities
22	Consumable Supplies
23	Methods and Procedures
24	Other
25	Other
26	Other

NO.	RK.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	0	10	5	10	25	20	5	20	5	0	20	5	10	20	10	5	20	10	0	40	20	10	10	20	0	0	0
2	1	35	4	10	12	3	18	5	8	0	35	1	10	15	6	20	5	8	0	60	15	10	5	10	0	0	0
3	1	10	10	20	40	20	0	0	0	0	10	10	40	30	10	0	0	0	0	20	10	10	10	50	0	0	0
4	2	10	45	9	5	5	20	5	1	0	30	10	20	15	5	5	10	5	0	25	5	5	5	60	0	0	0
5	2	30	3	9	20	5	5	2	3	23	15	0	25	15	10	5	5	0	25	55	20	0	3	22	0	0	0
6	2	25	25	10	5	25	0	10	0	0	35	15	15	0	35	0	0	0	0	0	0	0	0	100	0	0	0
7	2	20	20	10	20	10	0	10	10	0	10	20	15	10	5	0	30	10	0	80	0	0	0	20	0	0	0
8	2	10	0	20	35	25	0	0	1	9	5	0	40	40	5	0	0	0	10	90	0	10	0	0	0	0	0
9	2	35	10	5	5	2	0	13	30	0	30	15	20	2	2	1	10	20	0	5	15	0	0	80	0	0	0
10	2	20	15	20	5	15	5	10	10	0	20	15	20	15	10	5	5	10	0	10	20	5	5	30	30	0	0
11	2	18	15	18	9	10	14	10	6	0	24	5	19	5	10	10	17	10	0	40	40	5	5	10	0	0	0
12	2	10	65	10	2	3	0	2	8	0	8	70	7	3	0	0	2	10	0	5	10	0	5	80	0	0	0
13	2	5	30	20	5	5	0	35	0	0	1	35	28	1	0	0	35	0	0	5	10	30	35	20	0	0	0
14	2	28	22	17	20	7	2	2	2	0	35	14	20	10	10	1	5	5	0	20	25	15	3	37	0	0	0
15	3	12	35	7	10	30	1	0	5	0	15	20	10	10	30	5	0	10	0	-1	-1	-1	-1	-1	-1	-1	-1
16	3	10	3	50	0	30	0	5	2	0	10	10	40	0	35	0	3	2	0	5	45	0	45	5	0	0	0
17	3	15	15	40	5	5	0	10	5	5	10	10	40	10	10	5	5	5	5	5	40	10	5	40	0	0	0
18	3	20	15	10	10	5	0	10	0	30	30	10	10	5	0	0	10	10	25	3	25	25	2	45	0	0	0
19	3	25	5	15	10	10	10	20	5	0	20	5	15	15	10	5	25	5	0	40	10	10	10	30	0	0	0
20	3	20	20	25	5	10	0	0	5	15	10	35	5	5	5	0	0	5	35	10	5	15	5	30	35	0	0
21	3	10	20	20	5	20	5	10	10	0	5	5	10	5	10	5	30	30	0	10	10	10	5	65	0	0	0
22	3	30	40	10	10	5	2	2	1	0	50	30	10	10	0	0	0	0	0	10	0	0	10	30	50	0	0
23	3	15	17	18	15	15	11	7	2	0	20	15	15	15	15	12	7	1	0	30	30	10	15	15	0	0	0
24	3	0	75	5	5	0	0	0	15	0	0	80	0	0	0	0	0	20	0	0	0	0	0	100	0	0	0
25	3	10	0	10	60	15	0	0	5	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	80	5	0	0	15	0	0	0
26	3	20	5	30	10	5	0	0	30	0	25	5	25	5	5	5	0	30	0	10	20	10	10	50	0	0	0
27	3	9	41	9	23	9	4	3	2	0	10	30	20	10	15	10	5	0	0	30	5	2	3	60	0	0	0
28	3	15	60	7	5	4	3	1	5	0	20	60	3	10	5	1	0	1	0	5	0	0	0	95	0	0	0
29	3	22	8	4	38	8	8	6	6	0	25	5	5	35	5	10	5	10	0	84	2	2	2	10	0	0	0
30	3	5	15	20	5	5	10	5	30	5	5	5	35	5	5	5	5	30	5	10	5	5	5	65	10	0	0
31	3	35	10	15	10	10	5	2	0	13	18	15	15	7	7	7	8	15	8	50	5	1	5	39	0	0	0
32	3	40	10	10	25	5	5	5	0	0	40	4	20	20	5	1	9	1	0	30	30	15	10	15	0	0	0
33	3	15	5	20	20	10	10	15	5	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
34	3	50	10	30	0	0	0	0	10	0	50	10	30	0	0	0	0	10	0	1	1	1	1	96	0	0	0
35	3	10	20	25	5	20	0	10	10	0	15	15	15	5	25	0	15	10	0	5	30	3	2	50	10	0	0
36	3	10	10	25	15	10	5	15	10	0	25	10	20	25	5	0	10	5	0	50	15	15	10	10	0	0	0
37	3	20	15	10	10	20	0	10	15	0	20	5	16	10	20	5	12	12	0	-1	-1	-1	-1	-1	-1	-1	-1
38	3	20	15	20	0	1	14	20	10	0	10	10	20	0	0	0	20	40	0	90	0	0	0	10	0	0	0
39	3	20	10	30	0	10	0	20	10	0	30	0	30	0	10	0	30	0	0	30	0	30	0	40	0	0	0
40	3	3	5	20	50	10	10	0	2	0	40	10	10	20	10	5	0	5	0	40	20	15	15	10	0	0	0
41	3	10	30	20	5	20	14	1	0	0	25	25	20	10	10	5	5	0	0	30	30	10	0	30	0	0	0
42	3	10	5	25	30	10	5	10	5	0	25	3	20	25	10	2	10	5	0	75	5	5	0	15	0	0	0
43	3	15	2	60	4	2	0	13	4	0	30	1	49	5	2	0	10	3	0	50	5	30	0	15	0	0	0
44	3	20	10	20	25	10	15	0	0	0	20	10	5	10	30	20	5	0	0	40	20	5	10	25	0	0	0

45	3	10	20	20	10	5	5	30	0	0	20	20	20	10	5	5	20	0	0	10	10	10	30	10	30	0	0
46	3	30	10	5	50	5	0	0	0	0	20	5	5	60	10	0	0	0	0	75	0	5	5	15	0	0	0
47	3	15	5	15	3	4	1	5	5	47	-1	-1	-1	-1	-1	-1	-1	-1	-1	2	2	2	1	93	0	0	0
48	3	30	5	25	15	2	2	0	0	21	50	5	5	10	5	5	0	0	20	75	5	5	5	10	0	0	0
49	3	20	10	50	5	5	0	0	10	0	10	10	50	10	1	0	0	19	0	50	20	20	5	5	0	0	0
50	3	10	10	50	10	5	5	5	5	0	25	5	25	25	5	5	5	5	0	30	15	30	10	15	0	0	0
51	3	27	25	20	10	5	2	1	5	5	25	20	10	15	10	2	3	5	10	30	5	5	0	60	0	0	0
52	3	12	15	40	2	20	0	10	1	0	15	15	38	2	15	0	10	5	0	0	7	8	0	10	75	0	0
53	3	25	0	5	5	55	5	0	0	5	20	0	2	5	40	3	0	0	30	50	5	5	0	40	0	0	0
54	3	15	15	10	10	45	5	0	0	0	35	5	5	15	30	10	0	0	0	30	0	0	0	70	0	0	0
55	3	20	5	34	25	5	1	5	5	0	20	5	34	25	5	1	5	5	0	30	10	20	5	35	0	0	0
56	3	25	5	10	15	25	5	10	5	0	20	5	10	30	20	5	7	3	0	20	60	5	10	5	0	0	0
57	3	20	10	30	10	18	5	5	2	0	20	20	25	20	10	0	0	5	0	15	10	10	10	55	0	0	0
58	3	10	4	80	0	1	0	5	0	0	10	0	90	0	0	0	0	0	0	0	10	5	0	10	10	15	50
59	3	20	30	30	3	5	5.	5	2	0	20	20	20	5	5	10	15	5	0	15	20	5	5	55	0	0	0
60	3	11	9	10	2	20	28	8	12	0	20	5	8	0	22	22	5	18	0	99	0	0	0	1	0	0	0
61	3	20	5	10	30	20	5	5	5	0	20	1	5	50	10	5	1	8	0	70	15	2	3	10	0	0	0
62	3	5	20	15	5	50	0	0	5	0	5	10	40	5	30	0	0	10	0	10	5	0	5	80	0	0	0
63	3	10	30	10	30	5	2	3	10	0	5	50	5	30	5	0	0	5	0	50	0	0	20	30	0	0	0
64	3	15	10	15	10	10	10	30	0	0	10	5	15	10	5	5	50	0	0	10	25	25	25	15	0	0	0
65	3	5	20	20	35	10	5	2	3	0	10	5	15	60	5	1	4	0	0	60	14	5	1	20	0	0	0
66	3	10	10	25	20	25	3	2	5	0	5	0	25	5	5	20	20	20	0	65	15	3	2	15	0	0	0
67	3	5	10	60	10	3	5	5	2	0	10	10	50	23	2	2	1	2	0	70	10	10	5	5	0	0	0
68	4	15	10	15	15	5	25	10	5	0	20	5	15	15	5	25	10	5	0	50	5	10	5	30	0	0	0
69	4	15	20	10	10	5	1	2	2	35	10	20	15	2	0	0	3	10	40	25	15	9	1	50	0	0	0
70	4	5	15	40	25	5	10	0	0	0	10	10	45	25	5	5	0	0	0	90	0	10	0	0	0	0	0
71	4	15	10	22	25	20	2	5	1	0	25	5	20	15	25	4	5	1	0	50	25	1	4	20	0	0	0
72	4	10	10	10	30	20	10	5	5	0	10	20	20	10	10	20	5	5	0	20	15	15	10	40	0	0	0
73	4	16	12	21	33	3	4	4	2	0	15	5	15	30	10	15	5	5	0	70	3	8	1	18	0	0	0
74	4	30	20	10	0	20	0	10	10	0	30	10	20	0	15	0	10	15	0	70	10	5	5	10	0	0	0
75	4	30	25	20	2	1	0	10	12	0	50	15	10	1	1	0	13	10	0	-1	-1	-1	-1	-1	-1	-1	-1
76	4	10	10	25	10	10	10	20	5	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
77	4	38	10	20	1	10	1	10	10	0	40	5	23	1	5	1	20	5	0	1	1	1	2	95	0	0	0
78	4	15	5	30	20	15	10	5	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
79	4	15	5	25	10	10	5	15	15	0	20	5	20	8	5	2	20	20	0	30	10	25	0	10	25	0	0
80	4	15	10	25	10	10	10	15	5	0	20	10	30	10	5	5	10	10	0	15	2	7	1	75	0	0	0
81	4	35	2	20	15	15	8	3	2	0	35	5	20	20	15	2	2	1	0	80	10	5	2	3	0	0	0
82	4	50	10	10	5	5	0	15	5	0	15	10	25	5	10	0	25	10	0	35	10	10	5	40	0	0	0
83	4	30	5	30	5	15	5	5	5	0	30	5	30	5	15	5	5	5	0	40	10	0	0	50	0	0	0
84	4	5	10	5	50	20	6	2	2	0	15	5	5	45	20	5	0	5	0	40	30	10	5	15	0	0	0
85	4	20	15	25	15	15	10	0	0	0	20	15	25	15	15	10	0	0	0	20	15	15	10	40	0	0	0
86	4	8	12	30	18	3	2	7	8	12	18	4	25	19	3	1	8	8	14	40	2	3	15	40	0	0	0
87	4	5	10	70	2	5	0	5	3	0	6	12	58	3	12	0	6	3	0	40	2	0	1	57	0	0	0
88	4	15	20	25	10	10	5	5	10	0	50	20	10	10	5	5	0	0	0	15	25	10	0	50	0	0	0
89	4	15	15	20	20	15	5	5	5	0	10	10	20	40	5	5	5	5	0	45	5	5	15	30	0	0	0
90	4	10	20	25	5	10	15	13	2	0	20	10	30	5	10	10	15	0	0	25	10	25	5	35	0	0	0
91	4	20	20	5	30	16	7	1	1	0	25	5	10	38	15	5	1	1	0	45	15	15	5	20	0	0	0
92	4	0	0	50	0	20	0	20	10	0	0	0	50	0	10	0	35	5	0	5	0	0	20	75	0	0	0
93	4	20	20	40	0	8	0	10	2	0	40	10	40	0	0	0	10	0	0	0	1	0	0	99	0	0	0
94	4	25	15	40	0	0	0	20	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
95	4	30	10	5	5	5	1	4	5	35	40	10	2	2	1	2	1	2	40	95	1	1	1	2	0	0	0
96	5	10	5	25	20	10	5	20	5	0	10	5	15	15	10	10	30	5	0	49	1	0	1	49	0	0	0

4. Title of Billets

Index	Rank	Title of Billet	Other Responsibilities	Other Resources
1	0	Electronics Repair		
2	1			
3	1	Asst. Hull Repair Officer		
4	2			
5	2	Combat Information Center Officer, Electronic Warfare Officer	Watch standing 23/25	
6	2	Aviation Safety Officer		
7	2	Operational Squadron: Human Relations Officer		
8	2	Asst. Admin Officer	Report Prep 9/10	
9	2	Naval Gunfire Liaison Officer		
10	2	ADP Asst. System Director		
11	2	Staff Operations & Training Officer		
12	2	Analyst, Patrol Anti-Submarine Warfare Development Group COMFAIRWINGSLANT		
13	2	Contract Administrator		
14	2	Gunnery Officer, Destroyer		
15	3	Asst. Fighter-Attack Training Officer, COMFAIRNORFOLK		
16	3	Industrial Logistics Support Coordinator		

Index	Rank	Title of Billet	Other Responsibilities	Other Resources
17	3	Asst. VAQ Maint. Officer on COMMATVAQWINGPAC Staff	Staff Duty Off. 5/5	
18	3	Inertial Navigation (SSBN) Instructor Poseidon NAV Curric. Coordinator	Tech.Advisor 30/25	
19	3	Supply Officer		
20	3	Instructor	Teaching 15/35	Educat. Materials 35
21	3			
22	3	Asst. to the Comptroller, Budget & Plans Officer		Money 50
23	3			
24	3	Congressional Reporter/Liaison Officer of Legislative Affairs, Wash., D.C.		
25	3	Seamanship & Tactics Instructor		
26	3	Flag Lieutenant & Aide		
27	3	SPO Personnel Officer		
28	3	Annual Supply Officer		
29	3	Asst. Prof. of Naval Science, NROTC Recruiter for OK		
30	3	Weapons System Support Manager	Expeding High Priority 5/5	
31	3	Operations Officer	Watch Stand 13/8	
32	3	Surface Line, Operational Sea Tour as Department Head, Operations Officer		

Index	Rank	Title of Billet	Other Responsibilities	Other Resources
33	3	Weapons Officer DD		
34	3	Aide & Flag Lieut.		
35	3	Destroyed Squadron Material Officer		Repair Parts 10
36	3			
37	3	Navy Instructor		
38	3	Surface Warfare Assignment Officer (Detailer)		
39	3	Readiness Officer		
40	3	General Maintenance Div. Officer		
41	3	QA Division Officer, Patrol Squadron		
42	3	Supply Dept.-Control Billet		
43	3	Ship Superintendent Naval Shipyard		
44	3	Operational Avionics & Ordnance Division		
45	3	Stores Officer, USS Franklin D. Roosevelt		Non-consumable supplies 30
46	3	Instructor, Navy Supply Corps School		
47	3		Instructing	
48	3	Navigator (Flagship) Education & Training	Watch standing 21/20	
49	3	Aide to the Commander		
50	3	Asst. OPS Officer		
51	3		Briefings 5/10 Flying	

Index	Rank	Title of Billet	Other Responsibilities	Other Resources
52	3	Ship Supt. at LBNSY, Long Beach		Coord. Effort 75
53	3	Instructor/Observer FTG	Command Collat. Duty 5/30	
54	3			
55	3	Flag Lieutenant		
56	3	Engineer Officer		
57				
58	3			Repair Parts 50
59	3	ASW Advisor to Foreign Navy		
60	3	Aviation LCDR & Junior Officer (Detailer)		
61	3	Avionics/Armament Division Officer		
62	3	Undersea Warfare Analyst, Intelligence Div., CINCPACFLT Staff		
63	3	Admin. Officer		
64	3	Fiscal Supply Officer		
65	3	Line Division Officer		
66	3	Supervisor, CNO USN Plot		
67	3	Aviation Material Support Center Officer		
68	4	Director, Material Dept. XO, at Naval Supply Center, Newport, R.I.		

Index	Rank	Title of Billet	Other Responsibilities	Other Resources
69	4	Helicopter Anti-submarine Warfare Tactics Instructor	Instructing 35/40	
70	4	Branch Chief, Inventory Control Division at ICP		
71	4	Engineer Officer, USS Benjamin		
72	4	Chief Staff Officer, Coastal River Squadron One		
73	4			
74	4	Electronic Warfare Officer, Admin. Staff		
75	4	Supply Logistics Planning Officer		
76	4	Surface Junior Officer Detailer		
77	4	Project Officer, Plans & Policy Development Directorate		
78	4	Maintenance Officer		
79	4	Staff Civil Engineer		
80	4	Staff Billet Educational Development CNET		
81	4	Deck Officer CLG		
82	4	Electronic Warfare Officer		
83	4	Inventory Control Officer		
84	4			
85	4	Stock Control Officer		

Index	Rank	Title of Billet	Other Responsibilities	Other Resources
86	4	Operations Staff - Combined Navy-Air Force Operational Air Wing	Classified Intelligence 12/14	
87	4	Project Officer, Naval Air Systems Command Headquarters		
88	4	Director, Data Sys. & Analysis, Security Assistance Data Base		
89	4	Asst. Maintenance Officer		
90	4			
91	4	Flight Deck Officer		
92	4	Staff Oceanographer		
93	4	OPNAV - Foreign Asst. Div.		
94	4	Aircraft Integration Engineer & Coordinator		
95	4	Asst. NAVSECGRU Detail/ Placement Officer	Discussions 35/40	
96	5	Program Manager		

5. Estimation and Reallocation of Scores

Estimation:

Billet No.	Item: 19	20	21	22	23
15	75	5	5	0	15
33	10	40	0	20	30
37	50	0	5	0	45
75	5	5	10	75	5
76	90	0	5	0	5
78	25	55	5	5	10
94	25	30	5	0	40

Reallocation:

Billet No.	From	To
20	15% Teaching	15% Supervising
	35% Ed.Materials	35% Personnel
52	75% Coord.Effort	50% Personnel
		25% Facilities
69	35% Teaching	35% Supervising
95	35% Discussion	25% Negotiating
		10% Staffing

APPENDIX C

TEST APPLICATION OF HICLUST AND K-MEANS

A. HICLUST

In order to see how well Hiclust performs in identifying different groups in the data, it has been applied to a set of data with known structure. The well known Iris data has been selected because it has also been used as one set of test data by McRae [11] to validate his iterative K-Means algorithm. Thus, the findings of that study could be compared to the clustering performance of Hiclust.

There are three species of Iris flowers, Iris Setosa, Iris Versicolor, and Iris Virginica. From each species, the first 32 out of the original sample of 50 per species have been chosen as test data because Hiclust limits the number of objects to be less than 100. Iris Setosa are labeled 1-32, Iris Versicolor 33-64, and Iris Virginica 65-96. Four measurements have been collected on each Iris. They are: sepal length, sepal width, petal length, and petal width. It is known that on the basis of the four measurements the species Setosa is quite well separated from the two other species, whereas the Versicolors and the Virginicas overlap on some dimensions. Consequently, the samples from the last two species are more difficult to classify into the correct groups.

Ordinary euclidean distance has been selected as the dissimilarity measure. Two dendrograms have been obtained, one

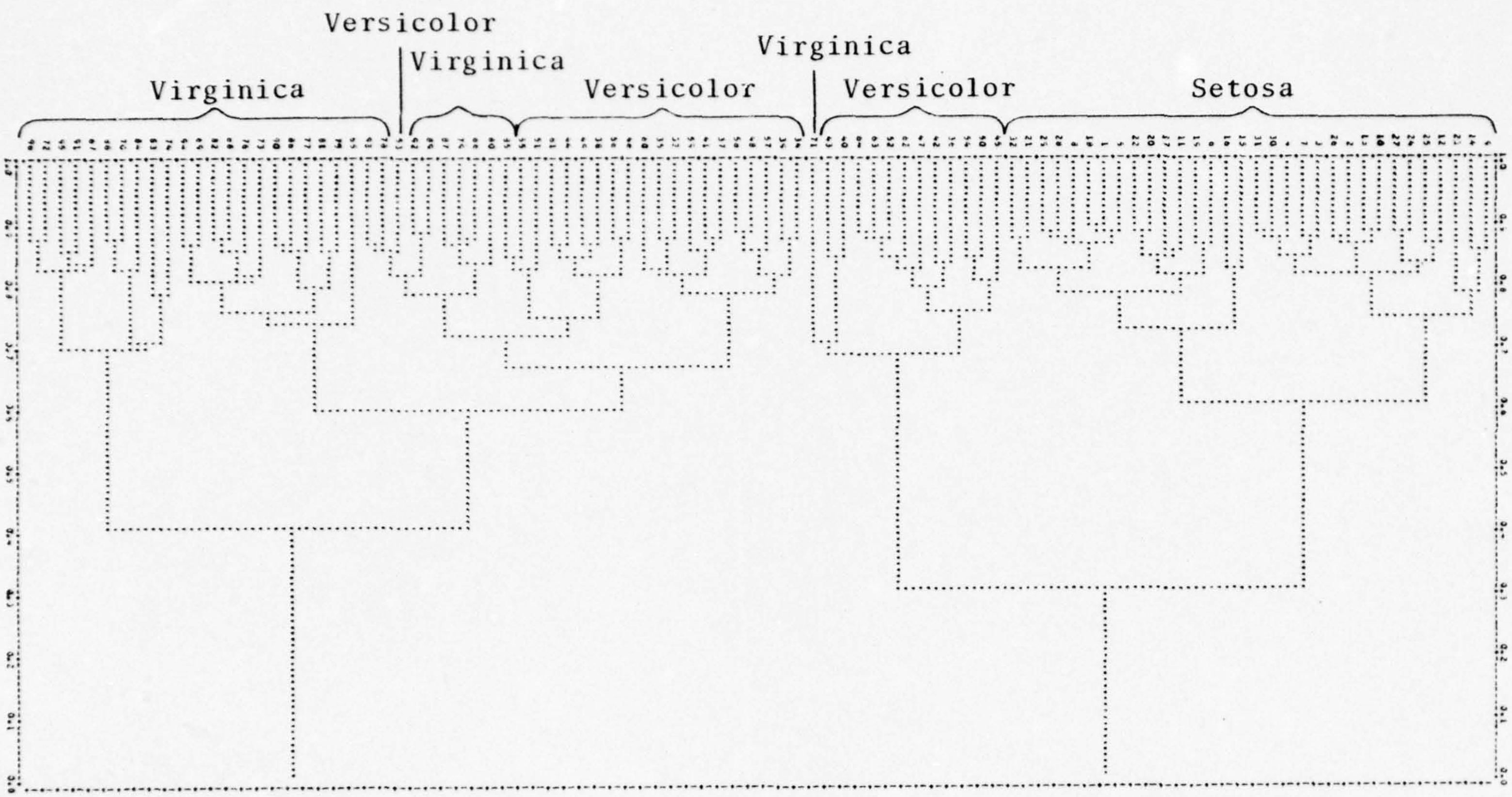


Figure 5. Dendrogram of Iris Flowers

for each option of algorithm (single link and complete link). The dendrogram generated by the complete link option is given in Figure 5.

At a level of .6, all Iris Setosa (Index 1 - 32) are grouped together in one cluster. The other clusters formed at that level contain a mixture of Versicolors and Virginicas. One can also observe that the dendrogram does not point to the presence of three groups. Similar results were observed by applying the single link option. Neither algorithm proved capable of separating the two overlapping species.

B. K-MEANS

The results of an extensive application of the K-Means program to the Iris data is contained in Ref. [11]. All 150 sampling units have been included (the program can process up to 600 objects and 20 variables). The number of misclassifications ranged from 3 to 25. The multivariate criteria, Det (W), largest root of $W^{-1}B$, and trace $(W^{-1}B)$, performed markedly better than the trace (W) criterion for whatever distance measure or standardization scheme had been chosen. The specification to standardize the data prior to the analysis always led to more misclassifications than without standardizing. The choice of the distance function had no effect on the number of misclassifications. It only influenced the rate of convergence.

A plot of the optimal Det(W) criterion vs. the number of clusters has pointed to the presence of three clusters.

The two-cluster solution groups together all Virginicas and Versicolors and separates the Setosas. The three-cluster solution then separates the first two species, also revealing the hierarchical relations.

APPENDIX D

CLUSTER SOLUTION

Cluster 1 Size: 5

No.	Rank	Billet Name
4	2	?
12	2	Analyst, Patrol ASW Development Group, COMFAIR WING SLANT
24	3	Congressional Reporter/Liaison Officer of Legislative Affairs
27	3	SPO Personnel Officer
28	3	Annual Supply Inspector

Cluster 2 Size: 29

No.	Rank	Billet Name
5	2	Combat Information Center Officer, Electronic Warfare Officer
7	2	Operational Squadron Human Relations Officer
8	2	Assistant Administration Officer
15	3	Assistant Fighter-Attack Training Officer, COMFAIRNORFOLK
25	3	Seamanship and Tactics Instructor
29	3	Assistant Professor of Naval Science NROTC Recruiter for OK
31	3	Operations Officer
37	3	Navy Instructor
38	3	Surface Warfare Assignment Officer (Detailer)
42	3	Supply Department - Control Billet

Cluster 2	Rank	Billet Name
#46	3	Instructor, Navy Supply Corps School
48	3	Navigator (Flagship) Education and Training
51	3	?
60	3	Aviation LCDR and Junior Officer (Detailer)
61	3	Avionics/Armament Division Officer
63	3	Administrative Officer
65	3	Line Division Officer
66	3	Supervisor, CNO USN Plot
67	3	Aviation Material Support Center Officer
70	4	Branch Chief, Inventory Control Division at ICP
71	4	Engineer Officer, USS Benjamin
73	4	?
74	4	Electronic Warfare Officer Administrative Staff
76	4	Surface Junior Officer Detailer
81	4	Deck Officer CLG
86	4	Operation Staff-Combined Navy-Air Force Operational Air Wing
89	4	Assistant Maintenance Officer
95	4	Assistant NAVSECGRU Detail/Placement Officer
96	5	Program Manager

Cluster 3 Size: 25

No.	Rank	Billet Name
1	0	Electronics Repair
2	1	?
3	1	Assistant Hull Repair Officer
14	2	Gunnery Officer, Destroyer
19	3	Supply Officer
20	3	Instructor
21	3	?
23	3	?
32	3	Surface Line, Operational Sea Tour as Department Head, Operations Officer
36	3	?
40	3	General Maintenance Division Officer
41	3	QA Division Officer, Patrol Squadron
44	3	Operational Avionics and Ordnance Division
49	3	Aide to the Commander
55	3	Flag Lieutenant
59	3	ASW Advisor to Foreign Navy
68	3	Director, Material Department, Executive Officer, at Naval Supply Center, Newport, R.I.
69	4	Helicopter Anti-submarine Warfare Tactics Instructor
72	4	Chief Staff Officer, Coastal River Squadron One
75	4	Supply Logistics Planning Officer

Cluster 3	Rank	Billet Name
#82	4	Electronic Warfare Officer
84	4	?
85	4	Stock Control Officer
88	4	Director, Data Systems and Analysis, Security Assistance Data Base
91	4	Flight Deck Officer
Cluster 4		**Size: 8**
13	2	Contract Administrator
18	3	Inertial Navigation (SSBN) Instructor, Poseidon NAV Curriculum Coordinator
39	3	Readiness Officer
43	3	Ship Superintendent, Naval Shipyard
50	3	Assistant OPS Officer
52	3	Ship Superintendent, LBNSY, Long Beach
64	3	Fiscal Supply Officer
90	4	?
Cluster 5		**Size: 3**
10	2	ADP Assistant System Director
45	3	Stores Officer, USS Franklin D. Roosevelt
79	4	Staff Civil Engineer
Cluster 6		**Size: 12**
6	2	Aviation Safety Officer
34	3	Aide and Flag Lieutenant

Cluster 6	Rank	Billet Name
#53	3	Instructor/Observer, FTB
54	3	?
57	3	?
62	3	Undersea Warfare Analyst, Intelligence Division, CINCPACFLT Staff
77	4	Project Officer, Plans and Policy Development Directorate
80	4	Staff Billet Educational Development, CNET
83	4	Inventory Control Officer
87	4	Project Officer, Naval Air Systems Command Headquarters
92	4	Staff Oceanographer
93	4	OPNAV-Foreign Assistance Division
Cluster 7		Size: 4
9	2	Naval Gunfire Liaison Officer
26	3	Flag Lieutenant and Aide
30	3	Weapons System Support Manager
35	3	Destroyed Squadron Material Officer
Cluster 8		Size: 7
11	2	Staff Operations and Training Officer
16	3	Industrial Logistics Support Coordinator
17	3	Assistant VAQ Maintenance Officer on COMMATVAQWINGPAC Staff
33	3	Weapons Officer DD
56	3	Engineer Officer

Cluster 8	Rank	Billet Name
#78	4	Maintenance Officer
94	4	Aircraft Integration Engineer and Coordinator

Table II

Centroids and Within-Cluster Standard Deviations (Normalized Scores)

Cluster (size)	Item 1	2	3	4	5	6	7	8	9	10	11	12	13
1 (5)	0.85	3.26	0.98	0.50	0.78	0.06	0.74	0.13	0.90	0.66	0.80	0.37	1.66
	0.54	1.09	0.15	0.64	0.33	1.46	0.24	0.94	0.50	0.33	0.26	0.24	0.70
2 (29)	0.02	0.23	0.21	0.47	0.03	0.00	0.16	0.07	1.05	0.54	0.53	0.30	0.59
	0.94	0.65	0.87	1.08	0.78	1.01	0.94	0.65	0.76	0.52	0.38	0.47	0.58
3 (25)	0.16	0.06	0.19	0.40	0.10	0.31	0.25	0.19	0.20	0.35	0.47	0.29	0.18
	1.08	0.60	0.73	1.08	0.52	1.11	0.80	0.63	0.52	0.63	0.46	1.33	0.62
4 (8)	0.40	0.07	0.73	0.67	0.37	0.23	1.19	0.45	0.45	0.03	2.45	0.29	0.37
	0.51	0.64	1.29	0.29	0.56	1.02	1.35	0.58	0.68	0.71	0.37	1.25	0.50
5 (3)	0.24	0.12	0.03	0.48	0.21	0.01	1.36	0.48	0.77	0.07	0.63	0.48	0.65
	0.49	0.59	0.21	0.22	0.50	0.00	1.34	1.27	0.42	0.45	1.25	1.51	0.43
6 (12)	0.31	0.28	0.46	0.77	0.96	0.43	0.14	0.10	0.75	0.74	0.73	0.30	1.49
	1.41	0.60	1.38	0.30	1.84	0.58	0.84	0.65	0.67	0.29	0.41	0.57	0.77
7 (4)	0.00	0.18	0.10	0.63	0.41	0.44	0.09	3.26	1.10	0.39	0.43	0.21	1.00
	1.30	0.50	0.80	0.19	0.81	0.87	0.73	1.67	0.11	0.81	0.51	0.41	0.53
8 (7)	0.01	0.45	0.62	0.36	0.15	0.09	0.38	0.36	0.70	2.48	0.46	0.60	0.53
	0.55	0.44	1.06	0.65	1.06	1.02	0.69	0.43	0.47	0.79	0.41	1.46	0.60

First row - Centroid

Second row - Standard Deviation

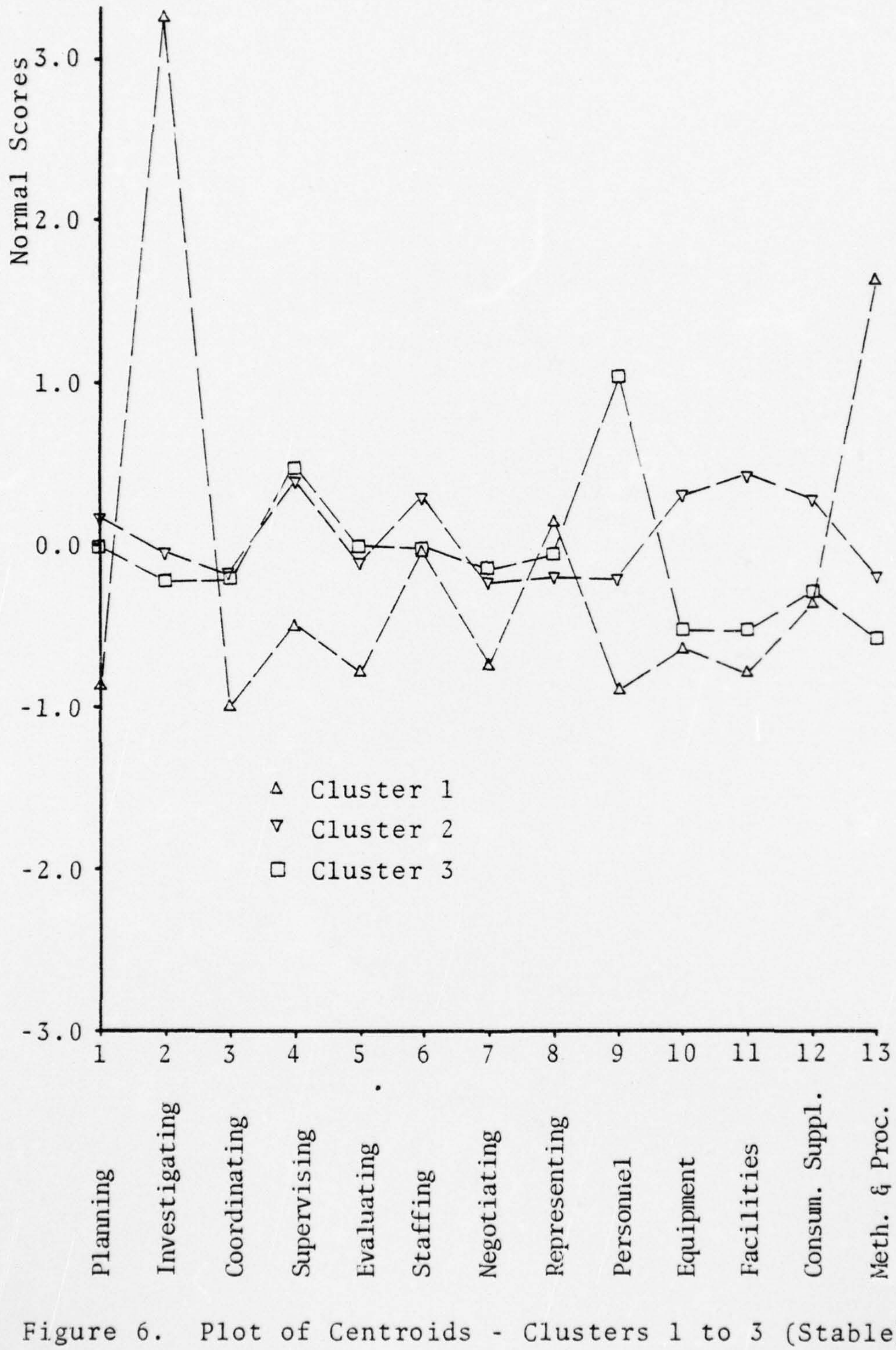


Figure 6. Plot of Centroids - Clusters 1 to 3 (Stable)

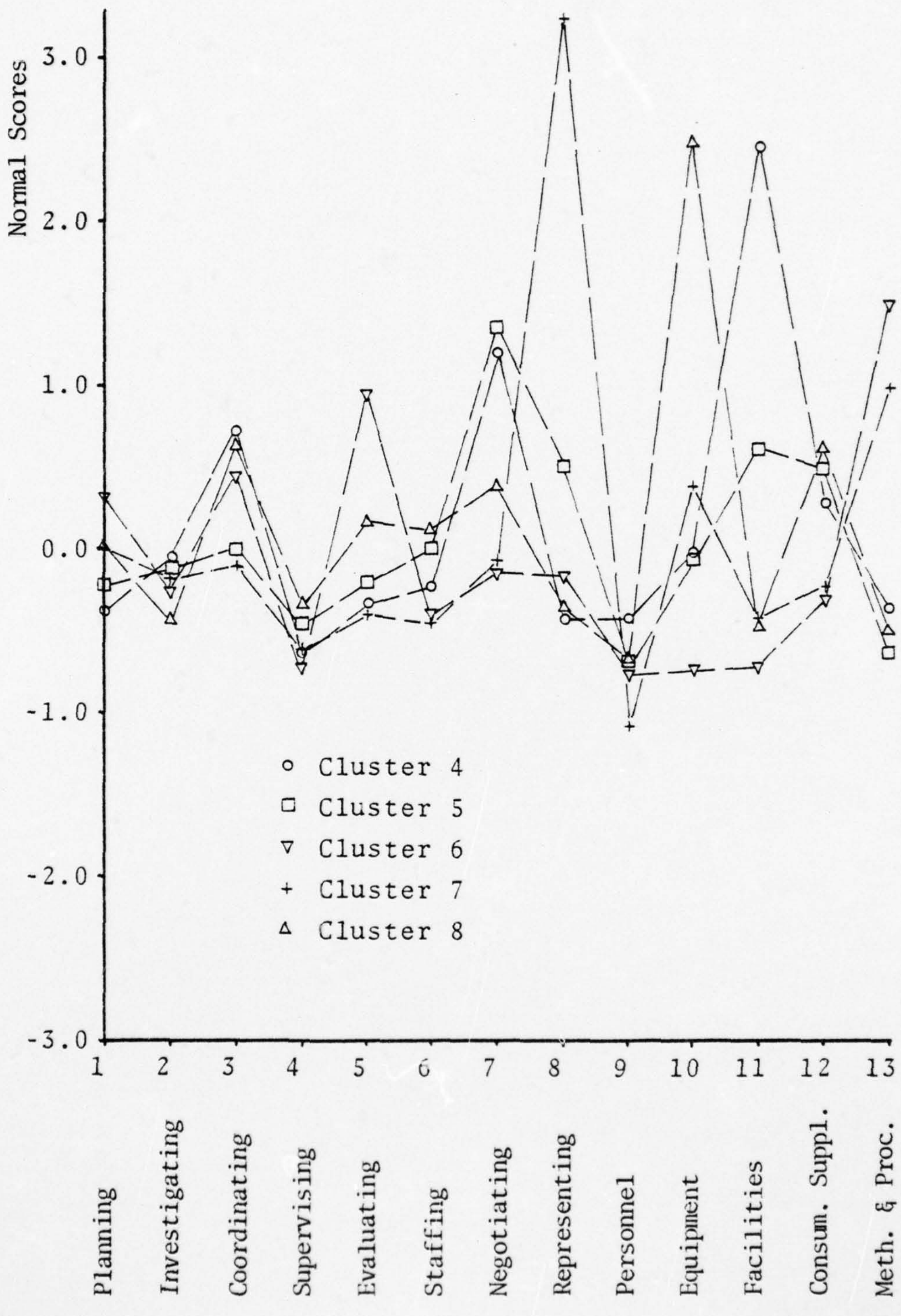


Figure 7. Plot of Centroids - Clusters 4 to 8 (Unstable)

Table III

Average Within-Group Variances of Normalized Scores Weighted by Cluster Size

Item	
Planning	.988
Investigating	.408
Coordinating	.882
Supervising	.752
Evaluating	.884
Staffing	1.000
Negotiating	.584
Representing	.557
Personnel	.389
Equipment	.333
Facilities	.214
Consumable Supplies	.968
Methods & Procedures	.378

APPENDIX E

RESULTS OF REPEATED GROUPING ANALYSIS (K-MEANS) WITH REDUCED DATA

The cluster solution shown in Table IV has been obtained by omitting all those sampling units for which missing scores have been estimated or reallocated (11 units), as documented in Appendix B. The numbers in parentheses designate the cluster index from the solution given in Appendix D. The removed responses are: 15, 20, 33, 37, 52, 69, 75, 76, 78, 94, 95.

Table IV

Cluster Solution (K-Means) Based on 82 Sampling Units

Cluster 1	2	3	4	5	6	7	8
4 (1)	5 (2)	1 (3)	16 (8)	9 (7)	10 (5)	3 (3)	53 (6)
12 (1)	7 (2)	2 (3)	56 (8)	26 (7)	13 (4)	6 (6)	54 (6)
24 (1)	8 (2)	11 (8)		30 (7)	18 (4)	17 (8)	62 (6)
27 (1)	25 (2)	14 (3)		34 (6)	39 (4)	21 (3)	
28 (1)	29 (2)	19 (3)		77 (6)	43 (4)	35 (7)	
41 (3)	31 (2)	23 (3)		92 (6)	45 (5)	55 (3)	
	38 (2)	32 (3)			50 (4)	57 (6)	
	42 (2)	36 (3)			64 (4)	72 (3)	
	46 (2)	40 (3)			79 (5)	80 (6)	
	48 (2)	44 (3)			90 (4)	85 (3)	
	60 (2)	49 (3)				88 (3)	
	61 (2)	51 (2)				93 (6)	
	63 (2)	59 (3)					
	65 (2)	68 (3)					
	66 (2)	82 (3)					
	67 (2)	83 (6)					
	70 (2)	84 (3)					
	71 (2)	86 (2)					
	73 (2)	87 (6)					
	74 (2)	89 (2)					
	81 (2)	91 (3)					
		96 (2)					

LIST OF REFERENCES

1. Occupational Research Center, Department of Psychological Sciences, Purdue University, Report No. 7, The Cluster Analysis of Jobs Based on Data from the Position Analysis Questionnaire (PAQ), by A.S. DeNisi and E.J. McCormick, September 1974.

2. Duran, B.S. and Odell, P.L., Cluster Analysis, Springer-Verlag, 1974, p. 1-31.

3. Friedmann, H.P. and Rubin, J., "On Some Invariant Criteria for Grouping Data," Journal of the American Statistical Association, 1967, 62 (320), p. 1159-1167.

4. Gower, J.C., "A General Coefficient of Similarity and Some of Its Properties," Biometrics, v. 27, December 1971, p. 857-871.

5. Hartigan, J.A., Clustering Algorithms, Wiley, 1975, p. 58-66.

6. Hemphill, J.K., Dimensions of Executive Positions: A Study of the Basic Characteristics of the Positions of Ninety-Three Business Executives (Bureau of Business Research Monograph No. 98), Columbus, Ohio: Bureau of Business Research, The Ohio State University, 1960.

7. Johnson, S.C., "Hierarchical Clustering Schemes," Psychometrika, v. 32, No. 3, September 1967, p. 241-254.

8. Kruskal, J.B., "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," Psychometrika, v. 29, No. 1, March 1964, p. 1-27.

9. Lance, G.N. and Williams, W.T., "A General Theory of Classificatory Sorting Strategies - 1. Hierarchical Systems," Computer Journal 9, No. 4, 1967, p. 373-379.

10. Mahoney, T.A., Jerdee, T.H. and Carroll, S.J., Development of Managerial Performance .. A Research Approach, Southwestern Publishing Co., January 1963, p. 1-67.

11. McRae, D.J., Clustering Multivariate Observations, Ph.D. Thesis, University of North Carolina, Chapel Hill, 1973.

12. Sokal, R.R. and Sneath, P.H., Principles of Numerical Taxonomy, Freeman, 1963.

13. Stogdill, R.M. and Shartle, C.L., Methods in the Study of Administrative Leadership, Bureau of Business Research 1955, p. 44-53

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California	2
3.	Department Chairman, Code 55 Department of Operations Research Naval Postgraduate School Monterey, California 93940	2
4.	Professor R. R. Read, Code 55Re (Thesis Advisor) Department of Operations Research Naval Postgraduate School Monterey, California 93940	2
5.	Associate Professor R. F. Boynton, Code 6412 Defense Resources Management Education Center Naval Postgraduate School Monterey, California 93940	1
6.	DOKCENTBW Friedrich-Ebert-Allee 34 53 Bonn, Federal Republic of Germany	1
7.	Luftwaffenamt Fliegerhorst 5 Köln 90 Federal Republic of Germany	1
8.	Professor D. Elster 7005 Southridge Dr. McLean, VA 22101	1
9.	Major Juergen Lemke, FGAF (Student) Schillerstr. 30 3062 Bückeburg, Federal Republic of Germany	1